NON-PARAMETRIC TESTING OF THE MACHINE LEARNING ELECTRICITY PRICES FORECASTS

ORIGINAL SCIENTIFIC PAPER / IZVORNI ZNANSTVENI RAD

UDK: 338.5:621.8.037

JEL: C53; Q47

DOI: 10.56321/IJMBS.10.16.5

Autor/Author:

DAVOR ZORIČIĆ
PhD, ASSOCIATE PROFESSOR
Faculty of Economics and Business, University of Zagreb
J.F. Kennedy Sq. 6, 10 000 Zagreb, Croatia
E-mail: dzoricic@efzg.hr
ORCID: 0000-0002-0206-3422

ABSTRACT

This research analyzes forecast accuracy in the day-ahead electricity market. Performance of Random Forest and XGBoost machine learning models is compared based on the day-ahead electricity market data for Germany. Data for 2018 and 2021 is analyzed in order to explore differences in forecast accuracy in the low and high market volatility periods. Initial training data for 2017 is used in order to produce forecasts for 2018 up to one month ahead. The training set is then rolled one month forward thus creating a fixed length rolling window of training and forecast set data for the remainder of the analyzed period. This methodological framework results in 11 forecasting sets for each analyzed year. Forecast accuracy is then evaluated by comparing root-mean-squared errors (RMSE) for the observed period. The focus of the research is on examination whether differences in the RMSE values of the competing machine learning models being analyzed can be reliably determined. For this purpose, firstly forecasting exercise has been conducted 30 times over for both machine learning models and each forecast set containing all forecast horizons. Secondly, median RMSE values are analyzed for each forecast set and non-parametric Wilcoxon rank-sum test is used to determine whether the observed differences in RMSE are statistically significant. Research results show small differences in RMSE values, however, they are found to be statistically significant for all forecast sets except one. Moreover, Random Forest seems to slightly outperform XGBoost model during the period of low market volatility, while XGBoost seems to perform better in the last three forecast sets of 2021 associated with higher market volatility.

KEY WORDS: forecast accuracy, day-ahead market, Wilcoxon rank-sum test, Random Forest, XGBoost, market volatility

1. INTRODUCTION

The global challenge to address the climate change issues has long moved the Renewable Energy Sources (RES) to the forefront of scientific research and funding support alike. As the investments in the RES continue to rise, new issues related to the inherent volatility of RES power production have to be addressed. This is increasingly putting under the spotlight research strands focused on the integration of the RES into the existing power system which are often, due to the mentioned volatility, exploring various energy storage systems options and aggregators as their likely operators. In this context both due to concerns related to operating complex power systems as well as their economic viability, electricity prices forecasting becomes one of crucial issues as presented in IRENA (2019). Therefore, as pointed out for instance by Weron (2014), electricity prices have become a key input in decision-making process of energy companies. The research such as Jurčević et al. (2022), Čović et al. (2021) or Braeuer et al. (2019) present examples of the important role the electricity prices forecasting plays in economic viability assessment of investment in energy storage facilities, without taking into account the classical interest stemming from conventional power production and trading activities.

Bearing the above mentioned in mind it is no wonder that the productivity of the electricity prices forecasting field has been overwhelming in the past decade. Multiple authors have therefore been dealing with reviews of the methods employed in order to provide classification of research efforts. Most notable examples include Weron (2014), Nowotarski and Weron (2018), Ziel and Steinert (2018), Cerjan et al. (2013) and Vlah Jerić (2020), with the last author focusing on

the statistical and artificial intelligence-based approaches in the review. The latter class of methods encompasses a broad subclass of artificial neural networks and the second subclass referring to other machine learning methods. The research in this paper focuses on electricity prices forecasting models belonging to this second subclass of methods. Out of many machine learning models listed in this category Support Vector Machine (SVM) or its extension Support Vector Regression (SVR) have been most widely employed, with Random Forest and XGBoost also being quite common according to research overview provided in Vlah Jerić (2020). Taking this into account performance of these models has frequently been compared. Notable research includes studies such as Lago et al. (2018), Zahid et al. (2019), Naumzik & Feuerriegel (2021), Didavi et al. (2021) and Jurčević et al. (2023).

This paper extends the mentioned research by delving deeper in the forecast accuracy examination. Given the reported results in Didavi et al. (2021) and Jurčević et al. (2023.), particularly the robust forecasting accuracy and algorithm running time performance of the Random Forest and XGBoost models found in the latter research, in this research the differences in forecasting errors of the two models are closely examined. Regarding the data and machine learning models forecasting methodology this paper draws heavily on the paper by Jurčević et al. (2023). However, in this study the forecasting model is rerun 30 times over for each forecast produced by both machine learning models in order to analyze whether this will affect the volatility of forecasts for the tested models and to determine whether there are statistically significant differences in their forecasting errors. Therefore, the main contribution of this paper to the existing literature is twofold. First, the conducted research aims to additionally test forecasting robustness of the analyzed models. Secondly, multiple forecasts produced are further used to determine whether there are statistically significant differences in the forecasting errors. In order to conduct the second part of the research Wilcoxon rank-sum test is used due to non-normality of the distribution of the obtained forecasts. Altogether, there are 22 forecasting sets spanning across the period of two years and covering both the period of low and increased market volatility in order to increase the validity of the results.

The paper is structured as follows. The second section contains description of data sampling methodology, machine learning methods used and the Wilcoxon rank-sum test. The third section presents the research findings and is followed by the conclusion.

2. DATA AND METHODOLOGY

2.1. Data, variables and sampling methodology

The wholesale electricity prices for the day-ahead market for Germany were collected from the ENTSO-E platform, along with the data on forecasts for wind and solar power generation, actual wind power generation, load forecasts and actual load and imbalance prices and volumes for years 2017, 2018 and 2021. EPEX-Database data was used to obtain electricity prices for the intraday market for 2017 and 2018. Simulation approach presented in Jurčević et al. (2022) provided the data for 2021. Machine learning models in the research utilize 24, 48, 168 hour lagged day-ahead prices and 24 hour lagged intraday price as independent variables. Also, 24 and 168 hour lagged actual wind generation are used. Lastly, dummy variables related to hour, day and month are employed alongside dummy variables for weekday, Saturday and Sunday.

The data for day-ahead prices is available in 60-minute intervals, while other variables have 15-minute frequency. Therefore, the dataset referring to the day-ahead prices was modified to match the 15-minute frequency by assigning the day-ahead electricity price data value associated with 60-minute interval to four associated 15-minute intervals.

Data for 2017 is used for training, cross-validation purposes and in order to produce for the first set of one month ahead forecasts for 2018. The training set is then rolled one month forward thus creating a fixed length rolling window of training, cross-validation and forecast set data for the remainder of the analyzed period. This methodological framework results in 11 forecasting sets for each year due to missing data which amounts to one training (and forecast) set in both 2018 and 2021. Data for 2018 and 2021 is analyzed in order to explore differences in forecast accuracy in the low and high market volatility since the market conditions in 2017 and 2018 were much more similar than in the 2021 as shown in Jurčević et al. (2022) and Jurčević et al. (2023).

With respect to data, variables and sampling methodology used, this research draws heavily on the paper by Jurčević et al. (2023). Therefore, the reader is referred to the mentioned paper for any details that may be omitted here in order to remain concise and focus on the novel elements in this study. These are, as already mentioned in the introduction section, related to stronger robustness testing and testing whether the differences in forecast accuracy are statistically significant. Details are presented in the next subsection.

2.2. Selected machine learning models and non-parametric testing

Based on the described data forecast accuracy is evaluated by comparing root-mean-squared errors (RMSE) for the observed period of the two competing machine learning models: Random Forest and XGBoost. Both methods are well known and described in various papers, e.g. Lago et al. (2018) or Zahid et al. (2019). Details related to the use of "caret" package in R as well as the "ranger" training method for Random Forest and "xgbTree" training method for XGBoost are again available in Jurčević et al. (2023). However, unlike in the mentioned paper in the case of which four forecasts were made for each forecast set, in this research 30 forecasts are made for each forecast set for both models in order to test whether this will affect the volatility of forecasts. Forecast accuracy is measured by relying on the root-mean-squared errors (RMSE) for each of 22 forecast sets.

Furthermore, median RMSE values are analyzed for each forecast set and non-parametric Wilcoxon rank-sum test is used to determine whether the observed differences in RMSE are statistically significant. Wilcoxon rank-sum test, first presented in Wilcoxon (1945), is used to test the hypothesis that the distribution of X measurements in the population A is the same as in population B which can be written as: $H_0 : A = B$. According to Wild and Seber (2000) the test is a non-parametric alternative to the two-sample t-test. Moreover, based on the same authors, when both sample sizes contain more than 10 observations the distribution of W_A can be treated as if it were Normal (μ_A, σ_A), where

$$\mu_A = \frac{n_A(n_A + n_B + 1)}{2} \tag{1}$$

and

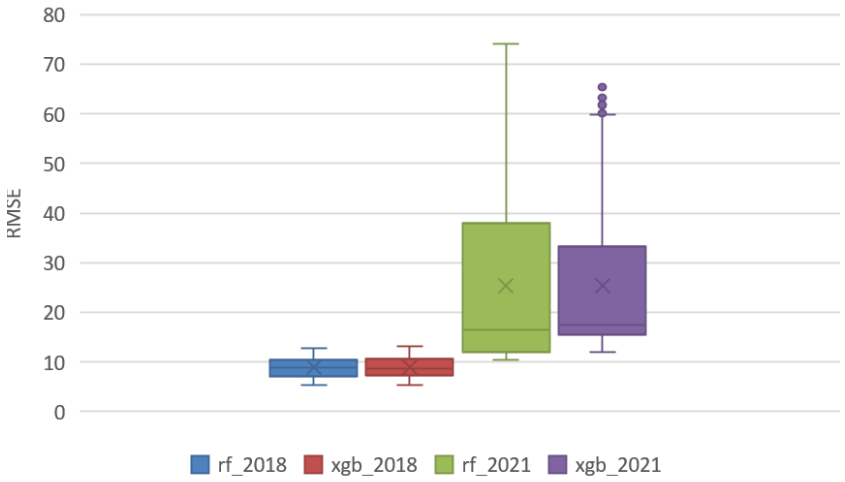
$$\sigma_A = \sqrt{\frac{n_A n_B(n_A + n_B + 1)}{12}} \tag{2}$$

with μ_A representing the sample mean, σ_A representing the sample standard deviation and n_A and n_B representing the A and B sample sizes respectively. Then probability of ($W_A \geq w_A$) approximately equals ($Z \geq z$), where $z = \frac{w_A - \mu_A}{\sigma_A}$ and $Z \sim$ Normal (0,1). It should also be noted that the test is still valid for any data distribution (not necessarily normal) and is much less sensitive to outliers than the two-sample t-test (Wild and Seber, 2000, p. 7).

3. FINDINGS

The research findings show that, as expected and already demonstrated in Jurčević et al. (2023), RMSEs are much smaller in 2018 than in the more volatile 2021. This is presented in the box plot chart (in Graph 1) which presents average RMSE of forecasts for all of the 11 forecast sets for each year and for which in this research forecasts have been carried out 30 times over. Regardless of the added complexity related to the new approach, results depicted in the box plot chart do not differ from the ones presented in the mentioned research even regarding the accuracy comparison of the two analyzed methods. Namely, Random Forest and XGBoost models are tightly matched in 2018 while in 2021, although the median and average RMSE of both models are similar, there is a bit more variation in the distribution related to the Random Forest model.

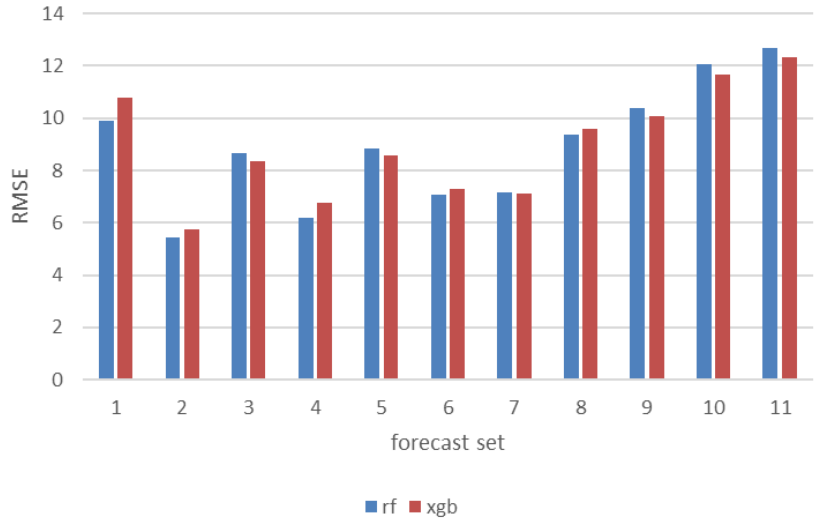
Graph 1. Box plot chart of average (RMSE) for analyzed machine learning models



Source: author's research

Accuracy of the analyzed models is further scrutinized by examining the differences in RMSE between the two models for each forecast set and analyzed year. The differences in median of the RMSE are used here because Wilcoxon rank-sum test, which is later employed, is carried out on median rather than the average values. Analysis for the 2018 (in Graph 2) provides further evidence of similar performance. Namely, out of 11 depicted forecast sets Random Forest model exhibited lower median RMSE value of forecasts in 5 instances as opposed to 6 in the case of XGBoost model. However, the biggest difference in the median RMSE values occurs in the first forecast set and in favor of the Random Forest model resulting in very small difference in median RMSE values overall.

Graph 2. Median of RMSE for each forecast set and analyzed machine learning model in 2018



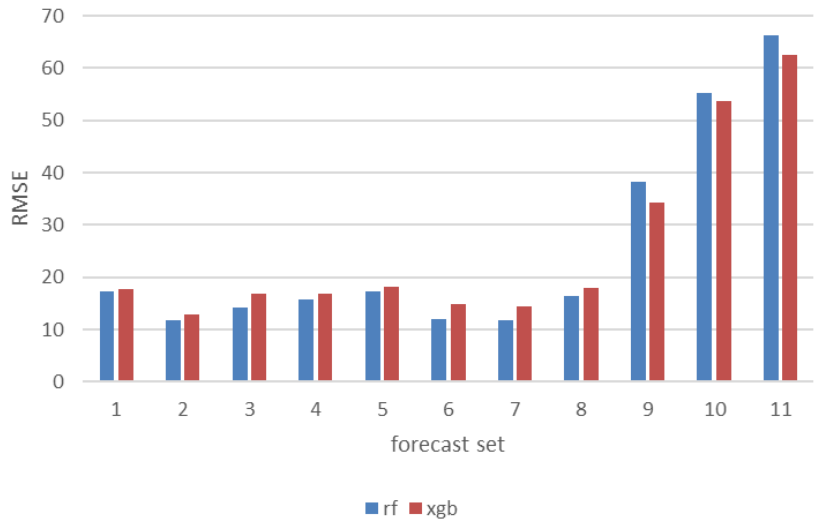
Source: author's research

In 2021 (depicted in Graph 3) the situation is a bit more complex. There is no pronounced difference in median RMSE values between models in the first 8 forecast sets, however, Random Forest slightly outperforms the XGBoost model.

In the last 3 forecast sets the situation is reversed with more pronounced differences in forecast accuracy in favor of the XGBoost model. It should also be noted that the last 3 forecast sets are characterized by a significant surge in RMSE values related to the increase in market volatility. However, overall in 2021 Random Forest outperforms the XGBoost model regardless of the hike in volatility which seems to deteriorate its forecast accuracy.

At this point it can be mentioned that, if forecast accuracy is ignored, XGBoost outperforms Random Forest model constantly in this analysis regarding algorithm running time by a relatively stable margin. On average the difference was just under 5 minutes in 2018 and a bit less at close to 4,4 minutes in 2021.

Graph 3. Median of RMSE for each forecast set and analyzed machine learning model in 2021



Source: author's research

Lastly, the test for statistical significance in differences in RMSE values is conducted. Due to the non-normality of the distribution of data related to 30 forecasts for each of the 11 forecast sets, non-parametric Wilcoxon rank-sum test is employed. The test tests for differences in median rather than averages (which would have been tested if the data distribution was normal and t-test was employed). Results of the test (in Table 1 below) show that only in the case of one forecast set in 2018 the differences in median RMSE values are not statistically different. For all other forecast sets hypothesis that the median RMSE values of compared data samples are the same is rejected at the 1% significance level as indicated by the p-values.

Table 1. Median of the Root Mean Squared Errors for each forecast set and analyzed machine learning models and p-value of the Wilcoxon rank-sum test

Forecast set	2018			2021		
	Random Forest	XGBoost	p-value	Random Forest	XGBoost	p-value
1	9,90	10,78	0,00000	17,35	17,73	0,00003
2	5,43	5,76	0,00000	11,81	12,79	0,00000
3	8,64	8,33	0,00042	14,28	16,94	0,00000
4	6,20	6,77	0,00000	15,69	16,88	0,00000
5	8,82	8,56	0,00000	17,27	18,26	0,00000
6	7,05	7,29	0,00000	11,96	14,74	0,00000
7	7,18	7,12	0,23985	11,83	14,36	0,00000
8	9,37	9,58	0,00350	16,44	18,00	0,00508
9	10,41	10,08	0,00000	38,28	34,27	0,00000
10	12,07	11,66	0,00000	55,22	53,62	0,00000
11	12,68	12,34	0,00009	66,19	62,45	0,00000

ISource: author's research

The results of the analysis conducted in this research show that the differences in RMSE values between the tested machine learning models are statistically significant, regardless of their seemingly small differences depicted in the charts. Overall, Random Forest model seems to perform slightly better, especially considering that one forecast set in 2018 in favor of XGBoost turns out not to be statistically significant. However, the research provides evidence supporting the view that in times of higher market volatility XGBoost model performs better in terms of accuracy. Coupled with its lower algorithm running time this seems to present a strong case in favor of the model in times of higher volatility.

4. CONCLUSION

The research analyzes forecast accuracy of electricity prices in the day-ahead market for two commonly used machine learning models. Research focus is on determining whether there is evidence of statistically significant differences in RMSE of produced forecasts. For that purpose, both models produced forecasts for each of 11 forecast sets in two analyzed years 30 times over. By conducting this exercise, the reported differences in RMSE do not seem pronounced. However, when Wilcoxon rank-sum test is employed, statistically significant difference in the median RMSE values is found in 21 out of 22 forecast sets. Moreover, even though the two models seem tightly matched, Random Forest model seems to perform slightly better than the XGBoost model in the period of low market volatility. During the high market volatility period the XGBoost model seems to yield lower RMSE values which is further supported by its lower algorithm running time.

The findings in this research complement and corroborate the findings in Jurčević et al. (2023) this study supports earlier findings comparing the forecast accuracy of electricity prices in the day-ahead market by Random Forest and XGBoost models. Namely, this research results, based on an increased number of forecasts provide evidence of forecast accuracy robustness of analyzed machine learning models. However, it also shows evidence supporting the view that the median RMSE values are statistically significantly different between the two models. Similarly Didavi et al. (2021) find XGBoost and Random Forest to outperform Decision Tree model by a wide margin. However, contrary to findings in this study XGBoost was found to clearly outperform Random Forest model.

Practical implications of this research can be evaluated in the context of the assessment of economic viability of investments in energy storage facilities and introduction of new market participants, such as aggregators in the electricity markets. To this end, relatively simple forecasting techniques producing robust forecasting results are warranted. Research findings in this paper and related studies offer valuable insights in this respect to the economic viability focused studies.

Paper limitations include narrow focus related to two analyzed machine learning models and the data sample based on only 2021 as more volatile year in the electricity market. Further research could benefit from using broader machine learning model base and more recent electricity market data considering the inherent market volatility associated with global geopolitical uncertainties.

LITERATURE

1. Braeuer, F., Rominger, J., McKenna, R. & Fichtner, W. (2019) Battery storage systems: An economic model-based analysis of parallel revenue streams and general implications for industry. Applied Energy, 239, issue C, pp. 1424-1440. DOI: 10.1016/j.apenergy.2019.01.050
2. Cerjan, M., Krželj, I., Vidak, M. & Delimar, M. (2013) A Literature Review with Statistical Analysis of Electricity Price Forecasting Methods. IEEE EuroCon 2013, pp. 756-763. DOI: 10.1109/EUROCON.2013.6625068
3. Čović, N., Braeuer, F., Mckenna, R. & Pandžić, H. (2021) Optimal PV and Battery Investment of Market-Participating Industry Facilities. IEEE Trans. Power Syst. 2021, 36(4), pp. 3441-3452. DOI: 10.1109/TPWRS.2020.3047260
4. Didavi, A. B., Agbokpanzo, R. G. & Agbomahena, M. (2021) Comparative study of Decision Tree, Random Forest and XGBoost performance in forecasting the power output of a photovoltaic system. 4th International conference on bio-engineering for smart technologies (BioSMART), IEEE, pp. 1-5. DOI: 10.1109/BioSMART54244.2021.9677566
5. IRENA. (2019) Innovation Landscape Brief: Aggregators. International Renewable Energy Agency; International Renewable Energy Agency: Abu Dhabi, United Arab Emirates
6. Jurčević, J., Pavić, I., Čović, N., Dolinar, D. & Zoričić, D. (2022) Estimation of Internal Rate of Return for Battery Storage Systems with Parallel Revenue Streams: Cycle-Cost vs. Multi-Objective Optimisation Approach. Energies, 15(16), p. 5859. DOI: 10.3390/en15165859
7. Jurčević, J., Vlah Jerić, S. & Zoričić, D. (2023) Electricity prices forecasting on the day-ahead market – performance comparison of selected machine learning models. In: Leko Šimić, M. (ed.) 12th International Scientific Symposium Region, Entrepreneurship, Development. Osijek, June 15-16, 2023. Osijek: Ekonomski fakultet Sveučilišta Josipa Jurja Strossmayera u Osijeku, pp. 591-602.
8. Lago, J., de Ridder, F. & de Schutter, B. (2018) Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy, 221, pp. 386-405. DOI: 10.1016/j.apenergy.2018.02.069
9. Naumzik, C. & Feuerriegel, S. (2021) Forecasting electricity prices with machine learning: predictor sensitivity. International Journal of Energy Sector Management, 15(1), pp. 157-172. DOI: 10.1108/IJESM-01-2020-0001
10. Nowotarski, J. & Weron, R. (2018) Recent advances in electricity price forecasting: A review of probabilistic forecasting. Renewable and Sustainable Energy Reviews, 81(1), pp. 1548-1568. DOI: 10.1016/j.rser.2017.05.234
11. Vlah Jerić, S. (2020). Statistical and artificial intelligence-based approaches to electricity price forecasting: A review. International Conference on Economics of Decoupling (ICED 2020). FEB Zagreb & Croatian Academy of Sciences and Arts, pp. 113-133.
12. Weron, R. (2014) Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting, 30(4), pp. 1030-1081. DOI: 10.1016/j.ijforecast.2014.08.008
13. Wilcoxon, F. (1945) Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), pp. 80-83. DOI: 10.2307/3001968.
14. Wild, C. J. & Seber, G. A. F. (2000) Chance Encounters: A first course in data analysis and inference, New York: Wiley.
15. Zahid, M., Ahmed, F., Javaid, N., Abbasi, R. A., Kazmi, H. S. Z., Javaid, A., Bilal, M., Akbar, M. & Ilahi, M. (2019) Electricity price and load forecasting using enhanced convolutional neural network and enhanced support vector regression in smart grids. Electronics, 8(2), 122. DOI: 10.3390/electronics8020122
16. Ziel, F. & Steinert, R. (2018) Probabilistic mid- and long-term electricity price forecasting. Renewable and Sustainable Energy Reviews, 94, pp. 251-266. DOI: 10.1016/j.rser.2018.05.038

NEPARAMETARSKO TESTIRANJE PROGNOZIRANJA CIJENA ELEKTRIČNE ENERGIJE STROJNIM UČENJEM

SAŽETAK

Istraživanje analizira prognostičku točnost na dan unaprijed tržištu električne energije. Uspoređuju se performanse "Random Forest" i XGBoost modela strojnog učenja temeljem podatka za njemačko dan unaprijed tržište električne energije. Podaci za 2018. i 2021. godinu analiziraju se kako bi se istražile razlike u prognostičkoj točnosti u razdobljima male i velike volatilnosti na tržištu. Inicijalni podaci za treniranje odnose se na 2017. godinu kako bi se napravile prognoze za 2018. godinu s prognostičkim horizontima do mjesec dana unaprijed. Uzorak podataka za treniranje zatim se pomiče mjesec dana unaprijed, čime se stvara pomični uzorak podataka fiksne duljine koji se koristi za treniranje i prognoziranje na ostatku analiziranog razdoblja. Ovakav metodološki okvir rezultira s 11 skupova prognoziranih podataka za svaku analiziranu godinu. Prognostička točnost zatim se ocjenjuje putem usporedbe korijena srednje kvadratne greške (engl. root-mean-squared error RMSE) u promatranom razdoblju. Fokus istraživanja je na ispitivanju mogućnosti pouzdanog utvrđivanja razlika u vrijednostima RMSE modela strojnog učenja koji se analiziraju. U tu svrhu najprije se opisani metodološki okvir za prognoziranje provodi 30 puta za oba modela strojnog učenja i za svaki skup prognoziranih vrijednosti sadržavajući sve prognostičke horizonte. Zatim se medijani RMSE vrijednosti analiziraju za svaki skup prognoziranih podataka te se provodi neparametarski Wilcoxonov test sume rangova kako bi se utvrdilo jesu li opažene razlike u vrijednostima RMSE statistički značajne. Rezultati istraživanja pokazuju male, ali statistički signifikantne, razlike u vrijednostima RMSE u svim skupovima prognoziranih podataka osim u jednom. Osim toga, čini se da Random Forest model rezultira nešto boljim prognozama od XGBoost modela u razdoblju niske volatilnosti. S druge strane, XGBoost model rezultira boljim prognozama u posljednja tri skupa prognoziranih podataka za 2021. godinu, a koji su povezani s povećanom volatilnošću na tržištu.

KLJUČNE RIJEČI: prognostička točnost, dan unaprijed tržište, Wilcoxonov test sume rangova, Random Forest, XGBoost, volatilnost tržišta