Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/136/2024.18817

Predicting China's Crude Oil Futures Prices: A Strategic Comparison of Random Forest and Time Series Models

Yunya Xia[1,a,*]

[1]*Faculty of Science and Technology, Beijing Normal University·Hong Kong Baptist University United International College, Jingtong Street, Zhuhai City, China*
a. r130018064@mail.uic.edu.cn
**corresponding author*

Abstract: This article examines the critical roles of Random Forest (RF) and Time Series (TS) models in forecasting China's crude oil futures prices, providing a comprehensive comparison of their predictive capabilities. The study employs ARIMA and SARIMA models, known for their proficiency in capturing data trends and seasonality, to harness the temporal aspects of oil price movements. In contrast, the RF model is recognized for its robustness in handling complex datasets, offering a nuanced approach to non-linear relationships and variable interactions. The analysis reveals that the ARIMA (0,1,4) model outperforms the (1,1,0) model in terms of prediction error and statistical fitting. However, the RF model's strength lies in its precision and flexibility, particularly in responding to market fluctuations. The paper concludes with the insight that ARIMA models are more suitable for long-term strategic planning due to their stability, whereas RF models excel in short-term forecasting and high-accuracy prediction scenarios. These findings are invaluable for market participants, offering them data-driven strategies to optimize their decision-making processes in the volatile oil futures market.

Keywords: Crude Oil Futures, Random Forest, Time Series Models, Forecasting Accuracy.

1. Introduction

The global energy market is inherently volatile and complex, with crude oil standing out as a key commodity that drives economic activities across the world. As the largest consumer of energy, China holds a significant position in the international crude oil market. A critical development in this market was the launch of China's crude oil futures on the Shanghai International Energy Exchange (INE) in 2018. This initiative provided a crucial platform for market participants to discover prices and manage risks effectively.

However, the unpredictability of oil prices presents a huge challenge. Factors such as economic indicators, and the dynamics of supply and demand contribute to this unpredictability, making it difficult for investors and policymakers to make informed decisions.

To navigate these challenges, predictive modeling has emerged as an indispensable tool. It enables stakeholders to anticipate market trends and make strategic decisions. Within the area of forecasting techniques, Random Forest (RF) and Time Series (TS) methods have become particularly prominent. RF, an ensemble learning technique, is praised for its robustness and its ability to manage large datasets with multiple features. It also provides insights into the importance of variables, which is

DOI: 10.54254/2754-1169/136/2024.18817

essential for understanding the factors that influence price movements [1]. Conversely, TS models, such as the Autoregressive Integrated Moving Average (ARIMA), excel at identifying trends, seasonality, and cyclical patterns within the data, making them valuable for time-based forecasting [2].

When employing time series and random forest models for option price forecasting, each model brings distinct developmental phases and characteristics that render them suitable for financial forecasting.

Early research typically used the ARIMA model in time series analysis, proposed by George E.P. Box and Gwilym M. Jenkins in the 1970s, which has become the cornerstone of time series prediction [3]. Initially, the ARIMA model was instrumental in managing univariate time series data through the inclusion of autoregressive (AR), integrated (I), and moving average (MA) elements. This framework revolutionized the capacity to predict based on the autocorrelation within data, a key aspect of forecasting. However, with the evolution of financial markets and an increase in data complexity, there arose a need for more advanced modeling techniques.

The ARIMA model's evolution gave rise to the Seasonal ARIMA (SARIMA) model, which addressed the original's shortcomings by factoring in seasonal patterns. This was especially beneficial for time series with periodic trends, such as those affected by quarterly financial reports or seasonal business activities. The SARIMA model enhanced forecasting accuracy by considering these seasonal variations [4].

Yet, the ARIMA model, like its predecessors, operates under the assumption of linearity, which may not be effective for non-linear data patterns. This has prompted the search for alternative models that can more accurately reflect the intricacies of financial markets.

In contrast, random forest models are an ensemble learning technique that functions by constructing multiple decision trees and consolidating their predictions to enhance accuracy and mitigate overfitting [1]. For option pricing, random forest models can manage non-linear relationships and interactions between variables, including implied volatility, interest rates, and the underlying asset price. They have proven particularly adept at capturing the intricate dynamics of options markets, where the price of an option is influenced by multiple factors. The robustness of random forest models in handling diverse data types, coupled with their capability to provide variable importance measures, makes them an appealing option for option pricing [1].

Despite its success, the random forest algorithm has faced challenges, such as the interpretation of the model, which is more complex than single decision trees due to the ensemble nature of the method.

In recent comparative studies, machine learning models, including random forests, have been pitted against traditional parametric models like the Black-Scholes model for option pricing [5]. The findings suggest that machine learning models surpass traditional models in terms of accuracy and robustness, especially when dealing with real-world data and market conditions that may deviate from the assumptions of parametric models [6]. Moreover, the application of machine learning techniques facilitates the inclusion of a broad spectrum of features, such as macroeconomic indicators and market sentiment, which can further bolster the predictive prowess of the models.

To begin with, historical data on China's crude oil futures prices must be analyzed and preprocessed to prepare a dataset suitable for modeling. Then, both Random Forest and Time Series models will be developed and calibrated using the prepared dataset. The performance of these models will then be evaluated using appropriate statistical metrics to validate their predictive accuracy. Finally, insights into the relative strengths and limitations of each method will be provided, along with recommendations for future research and practical applications in the energy sector.

This paper aims to contribute to the existing body of knowledge by offering a comparative analysis of these two prominent forecasting methods, thereby assisting market participants in making more accurate and data-driven decisions.

DOI: 10.54254/2754-1169/136/2024.18817

2. Method

In the context of Chinese crude oil futures prices, R language will be used to implement time series and random forest models for prediction (Figure 1).

Figure 1: Price trend chart of medium sulfur crude oil; Date from December 18, 2023 to September 20, 2024; Price in yuan/barrel.

2.1. Time Series Analysis

Firstly, crude oil price data was read and necessary data preprocessing was performed. The data processing steps include ensuring the dataset's integrity by removing any rows with missing values using the 'na.omit' function and calculating the logarithm of the price to stabilize the variance and achieve stationarity, a prerequisite for many time series models.

Subsequently, the Augmented Dickey-Fuller (ADF) test is conducted to check for stationarity in the time series. A stationary result (p-value < 0.05) confirms the effectiveness of differencing.

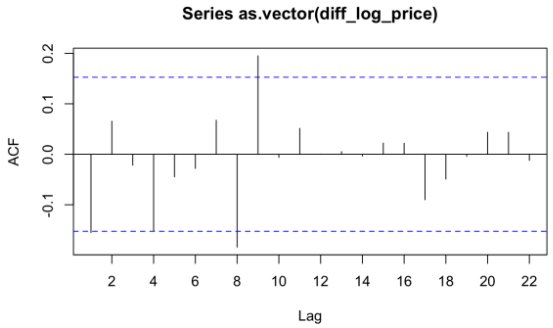

Figure 2: ACF.

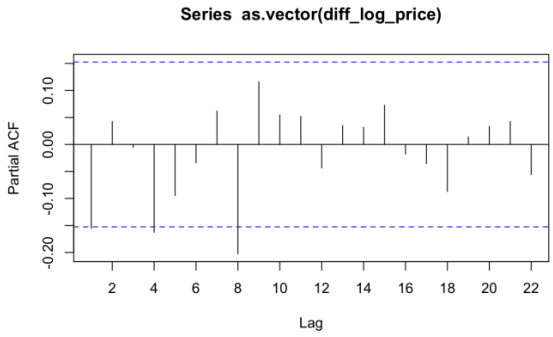

Figure 3: PACF.

DOI: 10.54254/2754-1169/136/2024.18817

The script then generates Autocorrelation Function (ACF) (Figure 2) and Partial Autocorrelation Function (PACF) (Figure 3) plots to identify potential lags significant for an ARIMA model, providing visual cues for selecting the appropriate model orders.

Following this, the 'armasubsets' function is utilized to identify potential ARMA models by fitting different orders of autoregressive moving average models. Four candidate ARIMA models are specified based on the ACF and PACF plots, and the Bayesian Information Criterion (BIC) is used to compare these models, with the lowest BIC indicating the best model fit that the ARIMA(1,1,0) model is optimal based on the BIC values (Table 1).

Table 1: BIC values.

Bic_values			
-860.5744	-865.4313	-851.4169	-854.9296

After model fitting, the residuals are analyzed to ensure they are randomly distributed, indicating no patterns that could suggest model misspecification. The script plots the standardized residuals, performs a normality check using a Q-Q plot (Figure 4), and examines the ACF and PACF of the residuals to ensure no significant autocorrelation remains.

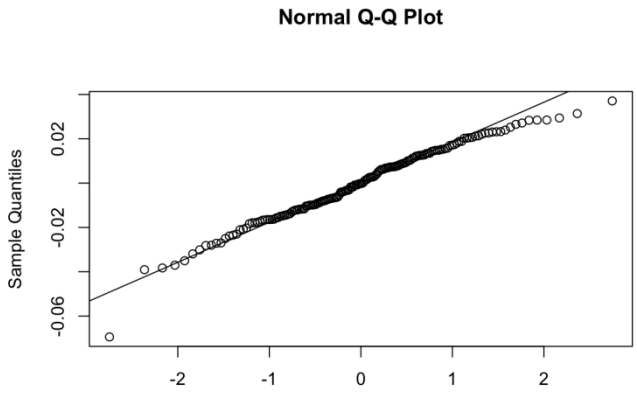

Figure 4: Q-Q plot.

Finally, the ‘forecast’ function is used to generate a forecast for the next 60 time periods based on the selected model. The forecast object is plotted to visualize the predicted values and confidence intervals, providing a comprehensive view of the future crude oil prices (Figure 5).

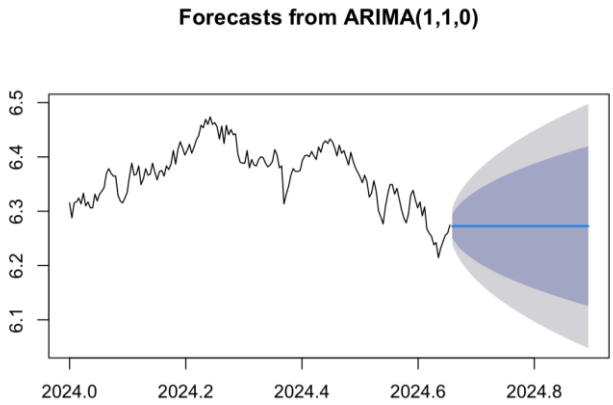

Figure 5: Time series prediction results (ARIMA).

DOI: 10.54254/2754-1169/136/2024.18817

By predicting the images, it was found that the price will remain at 6.8 yuan for the next two months, which is obviously not ideal. Considering that this set of data may have seasonal patterns, the SARIMA model was used to predict and obtain the ARIMA (0,1,4) model prediction results (Figure 6).

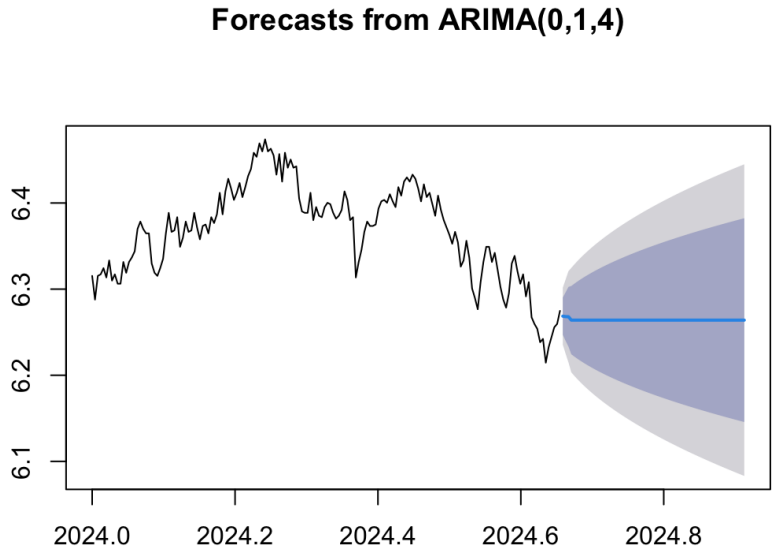

Figure 6: Time series prediction results(SARIMA).

2.2. Random Forest Algorithm

Firstly, crude oil price data was read and necessary data preprocessing was performed. The data processing steps include converting the date format to Date type and creating lagged features (such as lag1, lag2, lag3) to capture the characteristics of the time series.

Next, it will divide the data into a training set and a testing set, with July 1, 2024, as the boundary point. Then, the training data is fitted using a random forest model, which predicts crude oil prices through lagged variables. After completing the model training, it uses the data from the test set for prediction and calculates the root mean square error (RMSE) of the prediction to evaluate the model performance.

In addition, to visualize the prediction results, we calculated the standard error of the predicted values and constructed a confidence interval for the predicted values by doubling the standard error. The confidence interval provides a range of uncertainty for the predicted values, used to represent the credibility of the model's predictions.

In the visualization section, it used ggplot2 to draw a graph showing the actual values, predicted values, and confidence intervals. The actual values are represented by black solid lines, the predicted values are represented by blue dashed lines, and the confidence intervals are shown by blue shading. Finally, it added a legend to the chart, indicating the meanings of actual values, predicted values, and confidence intervals (Figure 7).

DOI: 10.54254/2754-1169/136/2024.18817

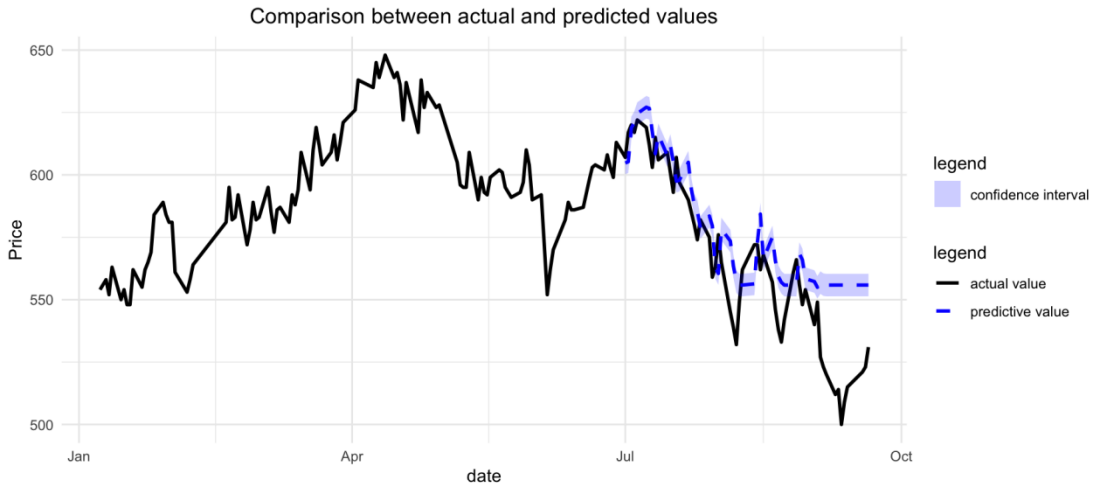

Figure 7: Random Forest Prediction Results.

3. Results Analysis and Case Study

3.1. ARIMA Model Forecasts

The ARIMA (0,1,4) model outperforms the ARIMA (1,1,0) model in multiple key indicators. The ARIMA (0,1,4) model has lower prediction error, better fit, and is statistically more favored (based on AIC and BIC) (Table 2 and Table 3). Therefore, it can be considered that the ARIMA (0,1,4) model performs better in these two models.

Table 2: ARIMA statistical indicators

sigma^2 = 0.0002922:	log likelihood = 437.82	
AIC=-871.64	AICc=-871.57	BIC=-865.43

Table 3: SARIMA statistical indicators

sigma^2 = 0.0002832:	log likelihood = 441.79	
AIC=-873.58	AICc=-873.21	BIC=-858.06

The coefficients of the SARIMA model, specifically the moving average terms (ma1, ma2, ma3, ma4), suggest the presence of certain autocorrelations in the series that the model has captured. The negative values of ma1 and ma4, along with the positive values of ma2 and ma3, indicate the complex dynamics within the price fluctuations. As shown in Table 5.

The ARIMA (0,1,4) (SARIMA) model has more parameters than the ARIMA (1,1,0) model (four moving average parameters compared to one autoregressive parameter), which may mean that ARIMA (0,1,4) can capture more complex dynamics in the data (Table 4 and Table 5).

Table 4: ARIMA Coefficient.

Coefficients:	
	ar1
	-0.1572
s.e.	0.0774

DOI: 10.54254/2754-1169/136/2024.18817

Table 5: SARIMA Coefficients.

Coefficients:				
	ma1	ma2	ma3	ma4
	-0.1593	0.0925	-0.0407	-0.231
s.e.	0.0775	0.0824	0.0779	0.0876

Choose the SARIMA (ARIMA 0,1,4) model to analyze the error metrics in its training set, including RMSE and MAE, which provide a quantitative evaluation of the model's accuracy. The RMSE of 0.01657402 and MAE of 0.0133827 suggest a reasonably good fit, with the model's predictions not deviating significantly from the actual values. The MPE and MAPE values further indicate that the model's errors are within acceptable limits, with no significant bias (as suggested by the ME value close to zero), as shown in Table 6.

Table 6: Rrror analysis.

ME	RMSE	MAE	MPE	MAPE	ACF1
-0.00042311	0.01657402	0.0133827	-0.007180828	0.2104078	-0.009376708

3.2. Random Forest Model Forecasts

The Random Forest model, as described in Table 7, offers a unique approach to predicting crude oil prices in China. Unlike the ARIMA model, which provides a smooth curve for price predictions, the Random Forest model's predictions show more variability, which is a reflection of its capability to capture complex, non-linear trends and interactions among multiple variables.

The model's predictions are dispersed around the actual price path, signifying a higher degree of variability in its forecasts. This feature is particularly beneficial for short-term predictions where market volatility is anticipated. However, it can also result in a greater degree of uncertainty.

As the mtry parameter increases, the model's performance improves, as evidenced by a significant decrease in the Root Mean Square Error (RMSE) and enhancements in the coefficient of determination (R-squared) and Mean Absolute Error (MAE). When the mtry value is set to 9, the model achieves its best performance, with the lowest RMSE, an almost perfect R-squared, and a very low MAE. These results suggest that the model not only has high predictive accuracy but also strong generalization capabilities.

The Random Forest model is particularly suited for scenarios that necessitate highly precise predictions. Its high R-squared value further indicates that it can effectively use input variables to account for changes in the target variable, which in this case is the crude oil price.

In summary, the Random Forest model performs optimally at a mtry value of 9 and is considered the preferred model for predictions due to its accuracy and robustness in handling complex data patterns.

Table 7: Training set error index.

mtry	RMSE	Required	MAE
2	132.25649	0.9954685	84.82225
5	80.602	0.9982557	38.99417
9	54.30226	0.9991988	18.4727

DOI: 10.54254/2754-1169/136/2024.18817

3.3. Comparative Strategy

Predictive performance: The RMSE of the ARIMA (0,1,4) model is 0.01657402, which is much lower than the RMSE of the random forest model at mtry=9 (54.30226). This indicates that the ARIMA model is superior in terms of prediction error, especially in measuring standard deviation.

Model complexity: The ARIMA model is relatively simple and mainly focuses on the autocorrelation of time series data. The random forest model is relatively complex, involving the integration of multiple decision trees, and can capture nonlinear relationships and interactions between variables.

Calculate cost: ARIMA models typically have lower computational costs, especially when the sample size is large. The random forest model has relatively high computational costs due to the need to construct multiple decision trees.

Application restrictions: The ARIMA model is sensitive to missing data and outliers.

Although the random forest model has some robustness to outliers, it may require a large amount of data to construct a stable prediction model.

Although the random forest model has a higher R-squared value at mtry=9, indicating its strong ability to explain the relationships between variables, the ARIMA (0,1,4) model is significantly better in terms of prediction accuracy (RMSE). Therefore, if the main focus is on prediction accuracy, the ARIMA (0,1,4) model may be a better choice. However, if the model needs to adapt to more complex data structures and capture nonlinear relationships, the random forest model is also a powerful tool after adjusting the appropriate mtry values.

4. Case Study: Predictive Performance

4.1. ARIMA Model Strengths and Limitations:

The ARIMA model excels in providing consistent trend predictions during periods of market stability, which is crucial for long-term planning. However, during times of market turbulence, the ARIMA model may not adapt quickly enough to sudden changes, potentially leading to less accurate short-term forecasts [7]. This is primarily because the ARIMA model relies on the autocorrelation of historical data, which may not be representative when market conditions undergo fundamental shifts.

4.2. Random Forest Model Strengths and Limitations

Conversely, the Random Forest model's responsiveness to market fluctuations is evident in its ability to predict short-term price movements with greater precision. Its ensemble nature allows it to adapt to new market information more rapidly, making it a valuable tool for tactical decision-making [8]. Random Forest improves accuracy and mitigates overfitting by constructing multiple decision trees and consolidating their predictions. Additionally, Random Forest can handle non-linear relationships and complex interactions between variables, which is particularly important in financial market forecasting.

4.3. Broader Application Value of the ARIMA Model

Despite the ARIMA model's potential limitations during times of market volatility, its predictive stability during relatively stable markets makes it an ideal choice for long-term planning. Moreover, the ARIMA model's parameter estimation and model-building process are relatively straightforward, making it easy to understand and implement, which is attractive for scenarios where rapid deployment of forecasting models is needed [9]. The ARIMA model can also be easily extended to Seasonal ARIMA (SARIMA) models to handle time series data with distinct seasonal patterns.

DOI: 10.54254/2754-1169/136/2024.18817

4.4. Broader Application Value of the Random Forest Model

The Random Forest model's ability to handle complex datasets and capture non-linear patterns gives it broad application potential in many fields beyond finance. For example, in healthcare, Random Forest can be used to predict disease progression and patient outcomes [10].

5. Conclusion

When comparing the ARIMA (0,1,4) model with the random forest model, it can find that the ARIMA model performs stably in time series data analysis, especially in long-term trend prediction, with intuitive and easy-to-understand parameters. It performs well when the data has obvious time series features, such as seasonal variations. However, ARIMA models have limited ability to handle nonlinear patterns and complex data structures and are sensitive to parameter selection, which may require domain experts to make adjustments.

In contrast, the random forest model demonstrated higher prediction accuracy in this case, with lower root mean square error (RMSE) and higher coefficient of determination (R-squared). This model is capable of handling nonlinear relationships and complex data structures, optimizing model performance by adjusting the maximum number of variables (mtry parameters) considered during decision tree splitting. The random forest model performs well in handling data with nonlinear features or complex interactions between variables, but compared to the ARIMA model, it is computationally more complex and resource-intensive.

Although the random forest model has advantages in prediction accuracy, its interpretability is not as intuitive as the ARIMA model. In addition, the random forest model has a certain robustness to outliers but may require a large amount of data to construct a stable prediction model.

Overall, ARIMA models are suitable for scenarios that require long-term planning and forecasting, while random forest models are suitable for short-term forecasting and situations that require high prediction accuracy. In practical applications, appropriate models can be selected or a mixed model approach can be considered based on the characteristics of the data and the prediction objectives. By combining the advantages of these two models, the accuracy and robustness of predictions can be improved.

It is important to note that while machine learning models offer significant advantages, they also present challenges. These include the need for large datasets for training, the risk of overfitting, and the complexity of model interpretation. Additionally, the stochastic nature of financial markets means that no model can guarantee perfect predictions and model performance should be continuously monitored and adjusted as market conditions evolve.

References

[1] *Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324*

[2] *Bontempi, G., Ben Taieb, S., & Le Borgne, Y. A. (2013). Machine learning strategies for time series forecasting. In *Lecture Notes in Business Information Processing* (Vol. 138, pp. 62-77). Springer. https://doi.org/10.1007/978-3-642-36318-4_3*

[3] *Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control* (Vol. 2). San Francisco: Holden-Day.*

[4] *Hyndman, R. J., Athanasopoulos, G., Razbash, S., Schmidt, D., Zhou, Z., Khan, Y., Bergmeir, C., & Wang, E. (2014). Forecasting functions for time series and linear models. In Package 'forecast'. Retrieved from https://cran.r-project.org/web/packages/forecast/forecast.pdf*

[5] *Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research*, 13, 1063-1095.*

[6] *Krauss, C., Fischer, T., & Huck, N. (2017). Deep learning with long short-term memory networks for financial market predictions. FAU Discussion Papers in Economics, No. 11/2017. Friedrich-Alexander-Universität Erlangen-Nürnberg, Institute for Economics. Retrieved from https://www.econstor.eu/bitstream/10419/157808/1/886576210.pdf*

DOI: 10.54254/2754-1169/136/2024.18817

[7] *Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(3), 1-22.*

[8] *Lin, Y., & Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474), 578-590. https://doi.org/10.1198/016214506000001398*

[9] *Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). New York: Springer.*

[10] *James, G., Witten, D., Hastie, T., & Tibshirani, C. (2013). An Introduction to Statistical Learning. New York: Springer.*