Forecasting crude oil prices based on machine learning statistics methods and random sparse Bayesian learning

Irina V. Pustokhina[*1], Denis A. Pustokhin[2]

[1] Department of Entrepreneurship and Logistics, Plekhanov Russian University of Economics, Moscow 117997, Russia
[2] Department of Logistics, State University of Management , Moscow 109542, Russia
Emails: Pustohina.IV@rea.ru; da_pustohin@guu.ru

Abstract

Oil price forecasting has received a great deal of interest from both professionals and scholars because of the unique characteristics of the oil price and its enormous impact on a wide range of economic sectors. In response to this problem, the authors set out to develop a strong model for accurately predicting the Brent crude oil price. We employed the Linear Regression and Random Forest models to examine the market interrelationships present in the oil price time series. Next, the models are given weights such that the experimental time series can be accurately predicted. These errors are quantified in terms of root mean squared errors (RMSE), average errors (MAE), and average percentage errors (MAPE). Results and forecast accuracy of the model as compared to the other model. To maximize their output and order levels and reduce the negative impact of potential shocks, countries that produce and import crude oil benefit greatly from accurate crude oil price forecasts.

Keywords: Linear Regression; Random Forest; Machine learning; Brent crude Oil; Forecasting.

1. Introduction

Financial markets around the world are greatly influenced by the price of crude oil. A solid estimate of oil prices and their variation has always attracted a lot of attention from investors and scholars working on oil markets[1], [2] [3]–[6] because of the impact that oil prices have on various financial markets (bond markets, share prices, etc.). Since they wish to be prepared for oil price variations, many nations' central banks and politicians appear to be keeping an eye on accurate oil price forecasts[7]. Because of this, it appears that accurate predictions of the crude prices are a necessity[8], [9].

Predicting and forecasting the oil price can be done using a variety of conceptual approaches. In the literature, there are two major classes of forecasting methods. GARCH, linear regression, error checking methods (ECM), autoregressive combined moving average (ARIMA), exponential smoothing model (ESM), and a random walk are all included in the first section of the course (RW). To the extent that these methods are beneficial for capturing linear relationships in time series, they overlook crude oil's nonlinear properties. It's been recommended that AI algorithms be used to solve this problem. ANNs, adaptive neuro-fuzzy inference systems, and smart optimization techniques like genetic algorithms have gained a lot of attention from scholars because of their strong learning capacities [10]–[13].

When it relates to time series models, the forecasting process is carried out utilizing past data. In this way, both linear and nonlinear models are used. Although it appears to be an arduous task to

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

determine if a correlation is linear or nonlinear, it is sometimes argued that a full linear or nonlinear relationship may rarely be distinguished [5] Consequently, and based on the relevant literature, no single forecasting method can lead to outstanding results under varied settings because each true issue has some specific complexity [14]. As a result, it is highly advised to use a combination of multiple models rather than just one[15]. It is possible that a hybrid model could be an effective medium for increasing the precision predictions. With the help of many forecasting models, it is easier than ever to see how time series are related to one another.

While statistical methods tend to perform well when applied to linear time series, they might struggle when dealing with nonlinear or nonstationary information. In order to do this, several different ML techniques were employed to forecast the cost of crude oil in current history. In principle, any regression approach may be used to anticipate crude oil prices since price forecasting is just another regression or prediction issue in ML. SVR and NN are two well-known ML methods. Crude oil price forecasts for the Brent and WTI (West Texas Medium) oil markets were made possible by Huang and Wang's integration of wavelet NN and random effective and convenient function to boost the performance of nonlinear estimate. The experimental findings confirmed that the method outperformed both BPNN and SVR. It was discovered by Xie et al. that SVR outperformed ARIMA and BPNN when used for forecasting. For oil price forecasting, Yu et al. employed GA to fine-tune the settings of an extended version of SVR called least squares SVR (LSSVR). The system was shown to be more efficient and accurate in making predictions than competing models. There are a number of other ML models utilised in this field as well, including hidden Markov models, MLRs, deep learning, and DFNPMs (data fluctuation nets predictive models).

2. Methodology

To foretell how much crude oil will be utilised, researchers turned to convolutional neural networks. The time series' properties were analysed statistically, and then the model was chosen. The statistical data was validated using the Augmentation Dickey–Fuller (ADF) experiment, which is a standardised unit root experiment, and its findings are interpreted by examining the p-value of such test. The null hypothesis (that there is no cointegration relationship and the sequence is stationary) is rejected if the statistic is less than or equal to 5%. In the event when p is more than 5%, the examined data series has an order of integration, is not stationary, and must be differentiated to become so. The conclusion of the study is provided in Table 4. There is insufficient evidence to rule out the null hypothesis since the p value for all conducted tests is greater than the chosen significance threshold of 5%. The time series that was examined did not pass the stationarity test.

2.1 Linear Regression

The first step in the data cleaning procedure is to eliminate any missing values from the dataset. Anaconda Jupyter Notebook tool is used to run a correlation study on the data set. It is then determined by using a Linear regression model, where Y is the overall number of Close prices of Brent oil per day, and X represents the total number of Open prices of Brent oil per day. The intercept and slope regression parameters are usually estimated using the least squares approach. An average peak value for Close prices is shown in Figure 1 below.

In Eq. (1), Y and X are the dependent and independent variables, s is the intercept, t is the regression parameter as slope, and n is the random error.

$$Y = s + tx + n(1)$$

Linear regression's drawbacks include the fact that it tends to focus on the mean of the input and output variables[16]–[18]. Linear regression is not a complete explanation of a single variable, just like the mean is not a complete description of a single variable. A Multiple Linear Regression (MLR) model is used to examine the numerous variables. Several independent variables influence the goal variable (the dependent variable). A regression equation that includes numerous variables can be described as such.

$$Y = s + t_1 z_1 + t_2 z_2 + t_3 z_3 + n(2)$$

The independent variables $z_1; z_2; z_3$ are known as the predictors or target variables. In this example, the y-intercept is called s and the coefficients are called $t_1; t_2; t_3$.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

2.2 Random Forest

Originally introduced by [19], [20], the random forest is an ensemble learning algorithm that builds several decision trees from various data subsets and votes on the findings of multiple decision trees to generate the random forest's output. Random forests have been demonstrated to be tolerant of outliers and disturbance, unlikely to overfit, and very accurate and stable predictors[1] [2].
For training, many unrelated decision tree types $[m(A,L); l = 1,\dots]$ are constructed using random forest. Every decision tree in the classification method makes a different forecast about the categorization of the sample. The mode of sample classification is the ultimate output. To reduce model variance, the random forest's performance can be enhanced by creating training sets that aren't related. Sample training yields $m_1(A)\dots m_l(A)$ sets of classifications, which are used to build the random forest model. Voting is used to decide the random forest's output, as shown in Equation (3).

$$M(a) = arg_z^{max} \sum_{i=1}^{L} D(m_i(a) = Z) (3)$$

The Random Forest model is represented by $M(a)$. $m_i(a)$ single decision tree model, with Z as the output and $(.)$ as the indicator function.
Each time a decision tree sampling is performed, a new training set is generated. There are N training subsets that are set up in playback that have a smaller number of training subsets than the whole number of training samples, and in most cases one-third of the total training samples.
Random forests are constructed from N decision trees. Determination trees are built for each training subset using the training set generated in the first phase. The CART method is used to divide nodes in a decision tree. Randomly selected objects are assigned to class I at node t based on the Gini coefficient reduction approach. p(j|t) represents the estimated likelihood that the objects in question are members of class j. The chance of misinterpretation is represented as Eq. 4 under this rule.

$$Gini = i \neq Jp(i|t)p(j|t) (4)$$

3. Experimental Datasets

On average, from January 2012 to July 2022, the daily Brent crude oil closing prices totalled 2650 dollars, as shown in Figure 1. We gathered this information. The average price of Brent crude oil in this time is $ 24.796849, the median is $22.2900925, and the standard deviation is $ 11.156475. Brent crude oil's erratic price movement underlines just how crucial it is to the market's forecasts. The highest price ($47.939999) was recorded on Feb. 24, 2012, as shown by the 690th observation in the time series. Another, the lowest price ($7.460000) was recorded.

Brent oil's price increased, according to the time series. International sanctions against Iran, one of the world's largest oil producers, and the occasional weakening of the US dollar have all contributed to an increase in oil demand in previous years. Between the middle of 2012 and the middle of 2014, the price fluctuated noticeably. Since the price of a barrel fluctuates between $45 and $50, we can classify this time as rather stable. Large economies including China, India, Russia, and Brazil are said to have reduced demand for oil between 2014 and 2016. Midway through 2016, prices began to rise and continued to do so into 2017. Brent oil's price is subject to several swings and fads. About 55% of the world's oil trade is benchmarked against the Brent oil price, making it an important factor in the global oil pricing system. Because of the wide range of patterns and surges that occurred throughout this period due to a variety of causes, using a model to analyze this period could lead to more accurate and dependable predictions. Then from 2022, the price began to increase due to the war in Ukraine.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

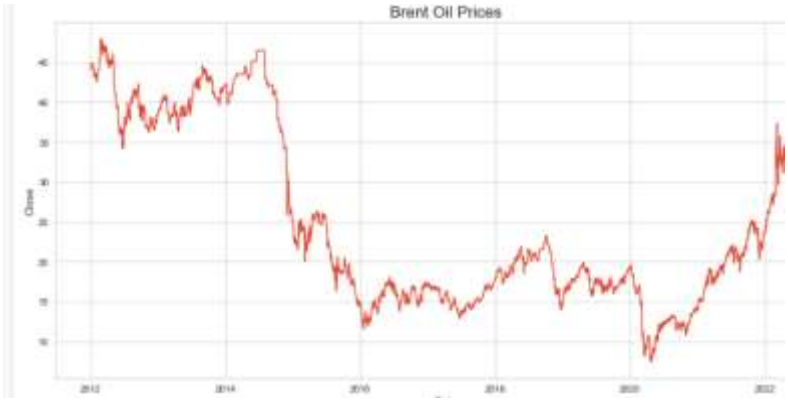

Figure 1: Daily Close price of Brent oil from 2012 to 2022.

Using time-series data, seasonality can be defined as a recurring pattern. While stock price fluctuations are predictable, they aren't the same as cyclical trends, which occur regularly but have no set length of time. Understanding your data's seasonality trends can provide a wealth of information and serve as a benchmark against which to assess your time-series machine learning algorithm. Figure 2 shows the seasonal and trend data. Figure 3 shows the pairwise relationships in a dataset. Figure 4 shows the heat map to represent information about the dataset with colors. Figure 5 shows the correlation between datasets by the correlation heat map. We make various statistics probability distribution methods in this study, then we take the largest ten methods and visualize them in figures 6 to 10.

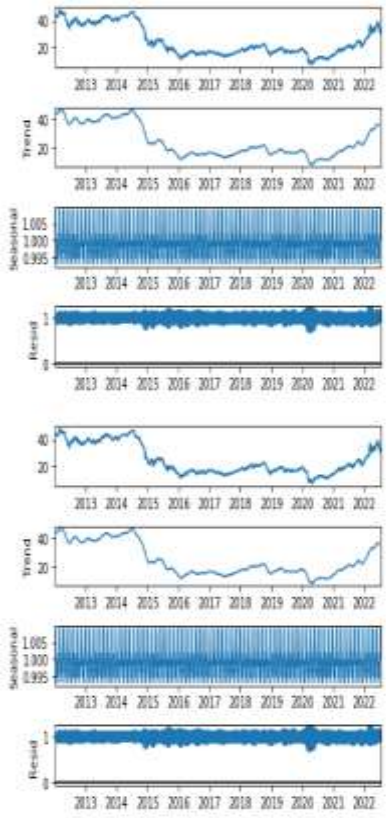

Figure 2: The seasonal and trend data.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

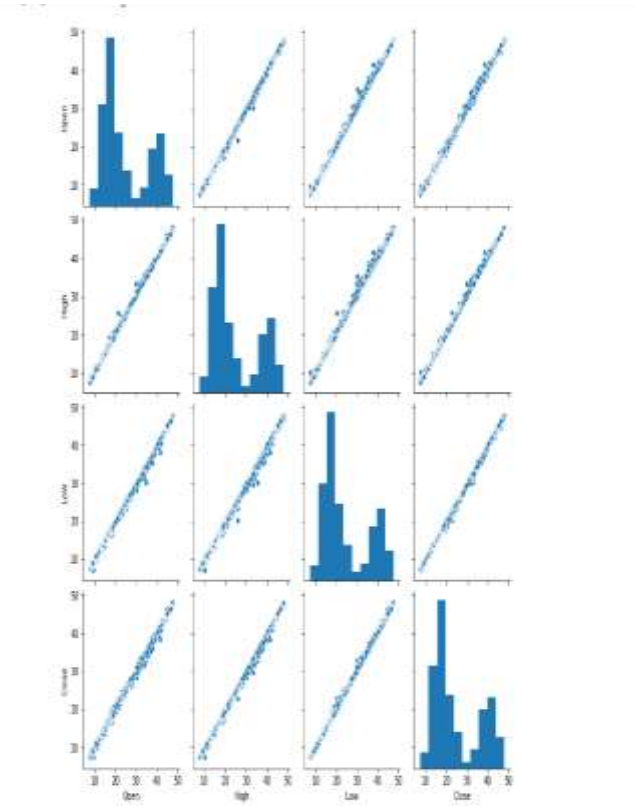

Figure 3: The pairwise relationships in a dataset.

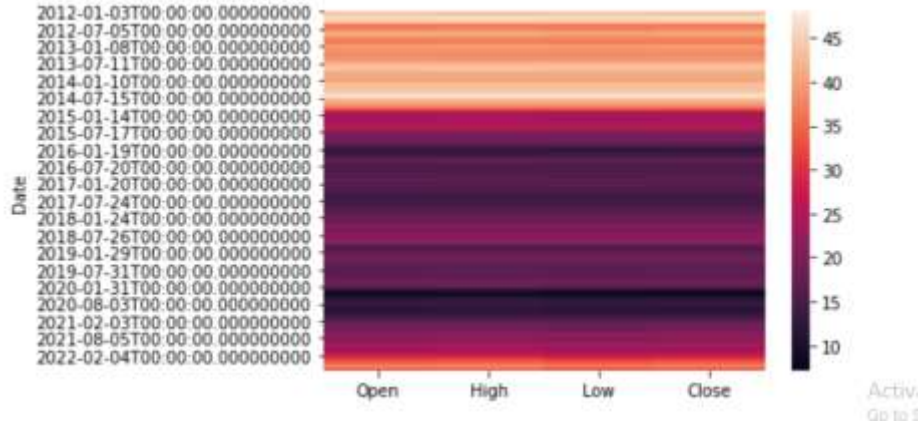

Figure 4: Visualized the information about the dataset in the heat map.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

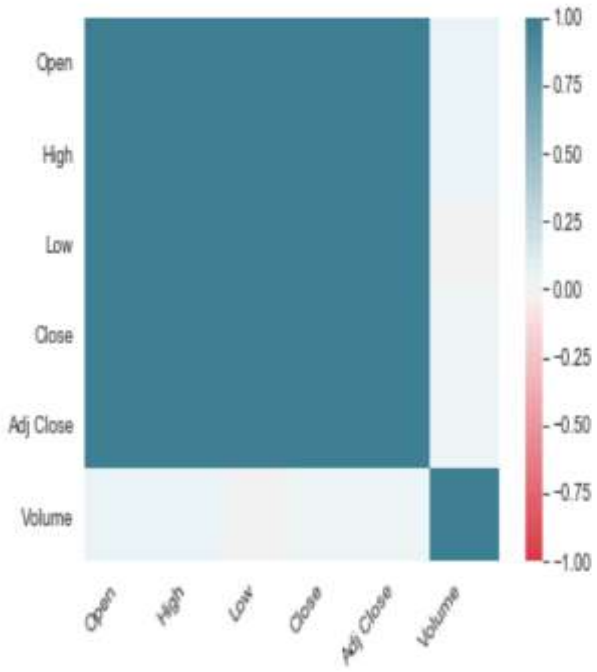

Figure 5: The correlation between datasets.

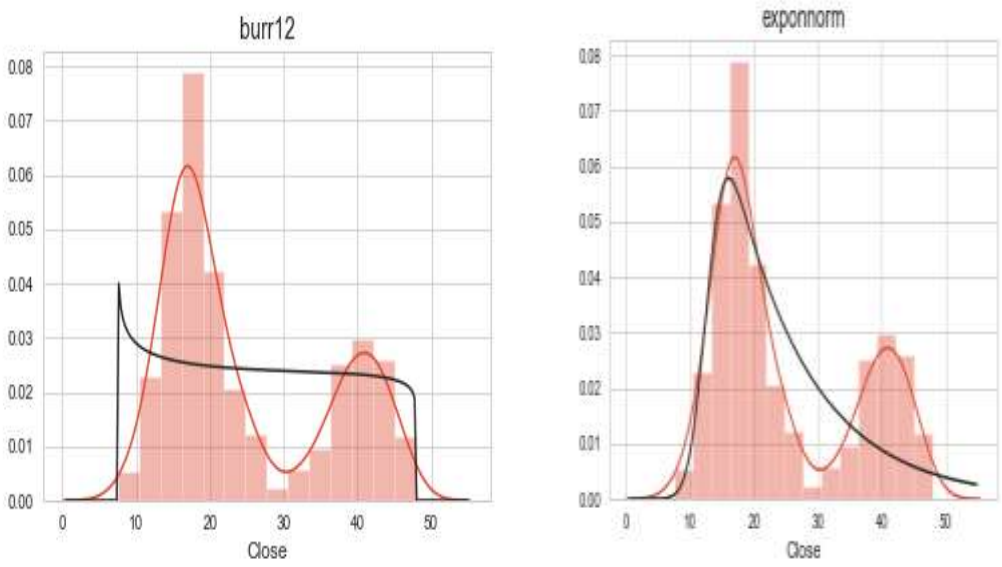

Figure 6: The statistics probability distribution burr12 and exponential normal.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

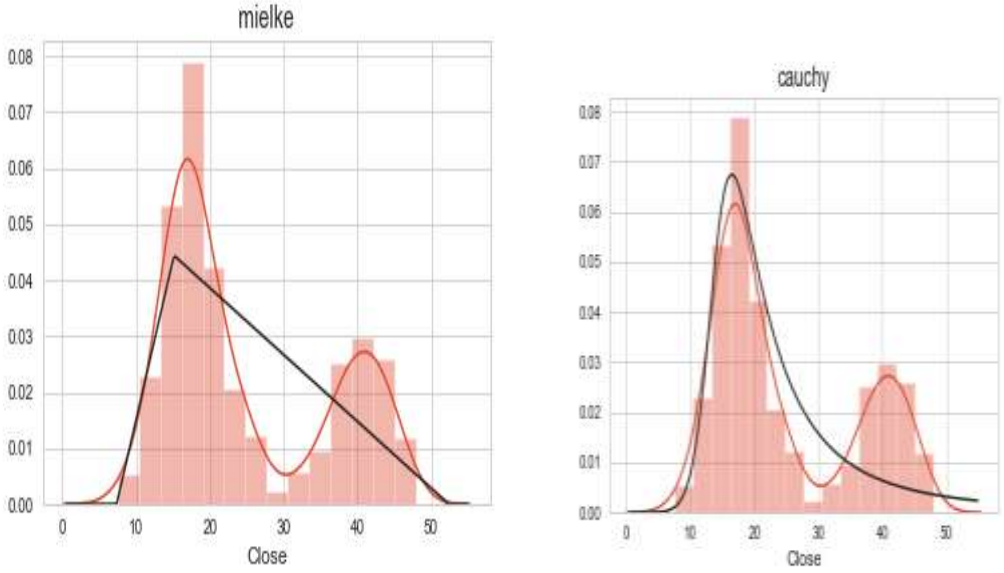

Figure 7: A statistics probability distribution for Mielke and Cauchy.

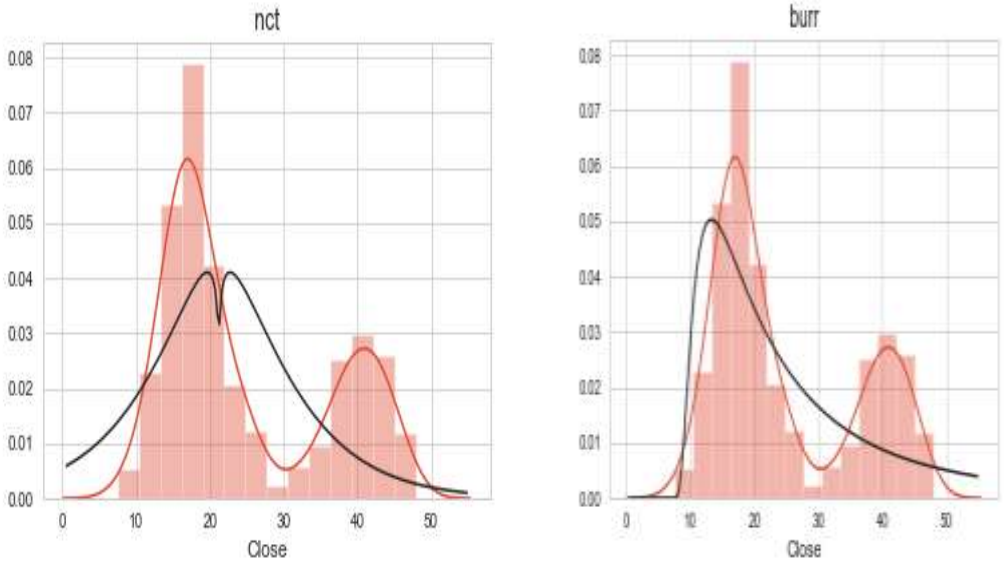

Figure 8: The statistics probability distribution for nct and burr.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

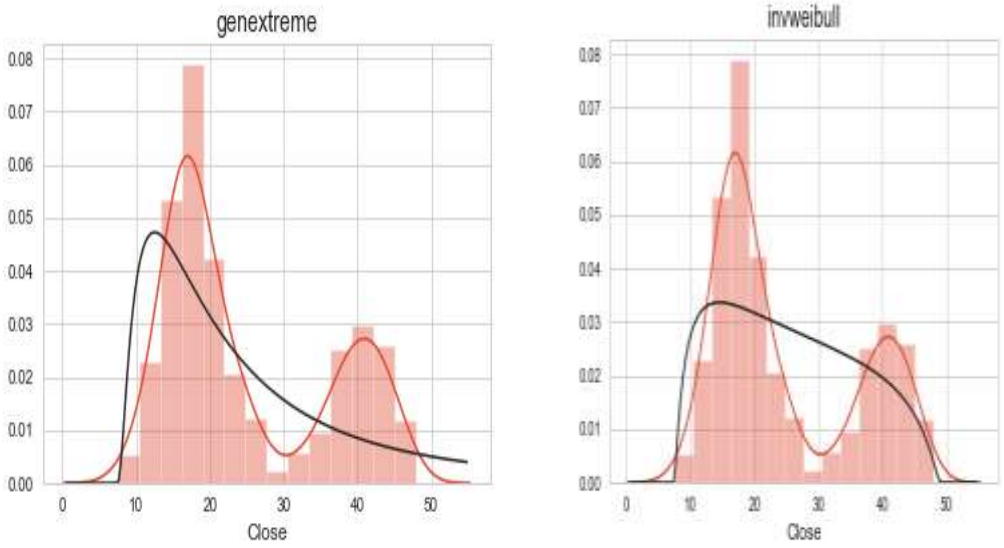

Figure 9: The statistics probability distribution for genextreme and invweibull.

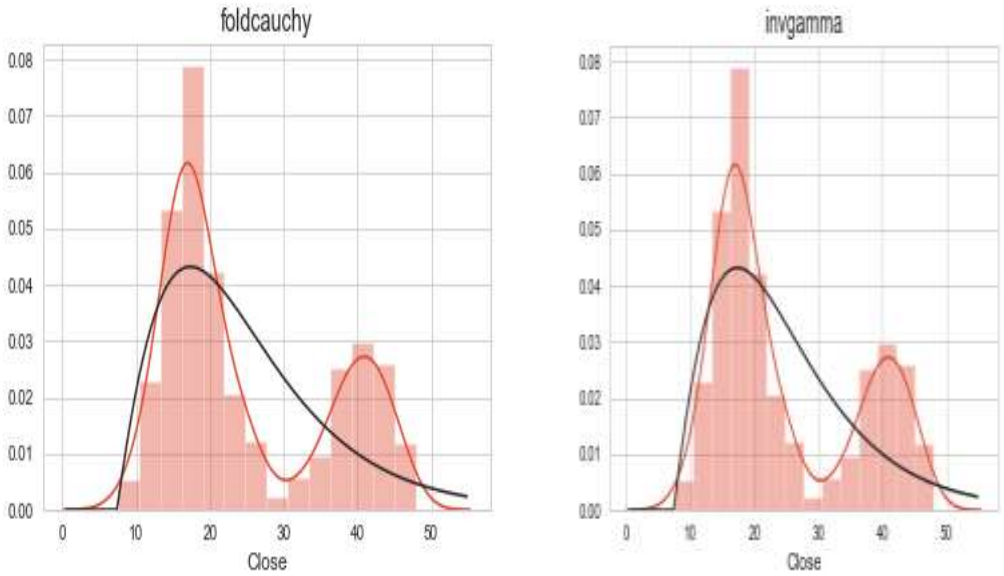

Figure 10: The statistics probability distribution for foldcathy and invgamma.

3.1 Results

Linear Regression is used to train 67% of the data set, which is then split into a training and testing set of 33% as shown in table 1. Table 2 shows the coefficients generated by the model to determine the test set's next price number. To summarise, the regression line finds the optimal intercept and slope values, resulting in a line that most closely matches the requirements.

In the context of linear multiple regression, regression with many variables is called linear multiple regression. To perform a multiple linear regression, you simply follow the same procedures as you would for a basic regression. There's a difference between the two when it comes to estimating. Use this to determine which factors have the greatest impact on anticipated output and how these elements are interconnected. Figure 11 shows the visualized actual and predicted values by the linear regression.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

We trained the random forest algorithm in the dataset as in table 1. The accuracy of random forest is 0.9333333333333333. The obtained accuracy by the linear regression is 0.99% and the random forest is 0.94%.

Table 1: Training and test values of dataset

	Percentage	Numbers
	2650	
Train	67%	1775
Test	33%	875

Table 2: The Linear Regression Values after training dataset.

Dataset	Values
Score	0.9993417227755259
Coefficient	0.99971304
Intercept	-1.8155683036980008e-06
Mean squared error	0.06761852707880976
Mean gamma deviance	0.00016781707204055506
Max error	1.951965301921355
R2 score	0.9994561913491461
Mean absolute error	0.1384087625734558
Mean squared log error	0.00014842802007420904

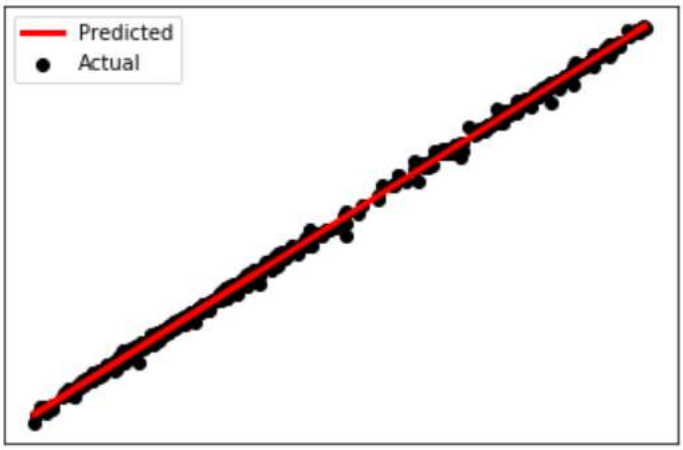

Figure 11: The visualization of actual and predicted values.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

3.2 Statistical Test

To further assess the AO-ANFIS even against contrasted theories, we use a well-known statistical technique, the Friedman test, below.
Table 3 shows the findings of the Friedman test. Tahe oilfields. First from tables, we recognize that the AO-ANFIS achieved the greatest results across the board (in terms of RMSE) with the exception of one oil hole.

Table 3: Results of Friedman test

Well No	AO	ANFIS	SMA	SCA	GWO
1	1.677	5.065	3.903	6.484	4.774
2	2.290	6.000	4.065	5.968	4.129
3	2.667	4.000	3.667	4.667	4.333
4	2.200	6.350	3.100	6.050	3.750
5	2.500	6.643	2.571	5.000	2.571
6	1.500	6.000	3.429	4.071	3.000
7	2.35	7.00	4.21	4.64	3.21

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

	7	0	4	3	4
8	1.571	6.571	3.1433	5.500	4.000
9	1.786	6.214	3.8577	5.000	3.000

Due to the country's steady and fast economic growth in recent years, crude oil consumption has skyrocketed in recent times. The oil reserves in Poland are insufficient to fulfil current needs. Crude oil usage has risen by 5.77 percent annually on average during the 1990s. The disparity between Poland's crude oil production and demand has been exacerbated by the country's increasing independence from foreign energy.

This article examines the geopolitical and financial underpinnings of Poland's energy raw goods market, crude oil supply, the composition of Poland's energy mix, and the hypotheses behind the country's energy strategy until 2040. The study results were utilised to inform an internal market model of raw oil consumption.

Polish statistics on yearly crude oil production from 1965 to 2020 were analysed. Table 4 displays descriptive data on the phenomena under study. Crude oil consumption in Poland is 18.51 Mtoe on average, which is quite near to the medium of 17.51 Mtoe. We find that the phenomena under study follows a platokurtic distribution. Every every piece of data has a positive skew.

Table 4: Some of statistics

	Measures
Mean	18.51
Standard error	0.92
Median	17.51
Standard deviation	6.74
Minimum	5.54
Maximum	32.82
Quantity	55.00
The largest	32.82
The	5.54

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

smallest	
Confidence level (95.0%)	1.84

4. Conclusion

As a result of the huge influence oil prices have on the global economy, many scholars have been compelled to produce accurate forecasts of this vital energy resource so that policymakers might plan for price shocks and limit their potentially disastrous economic implications. This complicates the analysis because relevant time series typically include lags and nonlinearities as well as connections to other markets. In the literature, providing a model capable of capturing and optimizing these aspects has been a problem. To forecast Brent crude oil prices, Linear regression and a random forest model were developed in this paper. Starting with the data collection, the authors compiled daily Brent oil price series from January 2012 to July 2022 before comparing the prediction results pairwise.

References

[1] M. Yang *et al.*, "Meta-analysis of acupuncture for relieving non-organic dyspeptic symptoms suggestive of diabetic gastroparesis," *BMC complementary and alternative medicine*, vol. 13, no. 1, pp. 1–12, 2013.

[2] D. Sun, H. Wen, D. Wang, and J. Xu, "A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm," *Geomorphology*, vol. 362, p. 107201, 2020.

[3] F. Picciolo, A. Papandreou, K. Hubacek, and F. Ruzzenenti, "How crude oil prices shape the global division of labor," *Applied Energy*, vol. 189, pp. 753–761, 2017.

[4] J.-Y. Wan and C.-W. Kao, "Interactions between oil and financial markets—Do conditions of financial stress matter?," *Energy Economics*, vol. 52, pp. 160–175, 2015.

[5] L.-T. Zhao, Y. Wang, S.-Q. Guo, and G.-R. Zeng, "A novel method based on numerical fitting for oil price trend forecasting," *Applied Energy*, vol. 220, pp. 154–163, 2018.

[6] R. B. Barsky and L. Kilian, "Oil and the macroeconomy since the 1970s," *Journal of Economic Perspectives*, vol. 18, no. 4, pp. 115–134, 2004.

[7] A. Safari and M. Davallou, "Oil price forecasting using a hybrid model," *Energy*, vol. 148, pp. 49–58, 2018.

[8] G. Wu and Y.-J. Zhang, "Does China factor matter? An econometric analysis of international crude oil prices," *Energy Policy*, vol. 72, pp. 78–86, 2014.

[9] C. Morana, "The oil price-macroeconomy relationship since the mid-1980s: A global perspective," *The Energy Journal*, vol. 34, no. 3, 2013.

[10] A. Azadeh, M. Moghaddam, M. Khakzad, and V. Ebrahimipour, "A flexible neural network-fuzzy mathematical programming algorithm for improvement of oil price estimation and forecasting," *Computers & Industrial Engineering*, vol. 62, no. 2, pp. 421–430, 2012.

[11] M. Hamdi and C. Aloui, "Forecasting crude oil price using artificial neural networks: a literature survey," *Econ Bull*, vol. 3, no. 2, pp. 1339–1359, 2015.

[12] S. Moshiri and F. Foroutan, "Forecasting nonlinear crude oil futures prices," *The energy journal*, vol. 27, no. 4, 2006.

[13] L. Yu, W. Dai, and L. Tang, "A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting," *Engineering Applications of Artificial Intelligence*, vol. 47, pp. 110–121, 2016.

[14] M. Khashei and M. Bijari, "A new hybrid methodology for nonlinear time series forecasting," *Modelling and Simulation in Engineering*, vol. 2011, 2011.

[15] A. Timmermann, "Forecast combinations," *Handbook of economic forecasting*, vol. 1, pp. 135–196,

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022

2006.
[16] T. K. Saha, S. Pal, and R. Sarkar, "Prediction of wetland area and depth using linear regression model and artificial neural network based cellular automata," *Ecological Informatics*, vol. 62, p. 101272, 2021.
[17] A. Sultan, W. Sałabun, S. Faizi, and M. Ismail, "Hesitant Fuzzy linear regression model for decision making," *Symmetry*, vol. 13, no. 10, p. 1846, 2021.
[18] F. M. Ottaviani and A. De Marco, "Multiple Linear Regression Model for Improved Project Cost Forecasting," *Procedia Computer Science*, vol. 196, pp. 808–815, 2022.
[19] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

DOI: https://doi.org/10.54216/AJBOR.070202
Received: March 06, 2022 Accepted: August 20, 2022