MACHINE LEARNING MODELS TO PREDICTION OPEC CRUDE OIL PRODUCTION

by

Hiyam ABDULRAHIM[a*], Safiya Mukhtar ALSHIBANI[b], Omer Ibrahim Osman IBRAHIM[c], and Azhari A. ELHAG[d]

[a]Department of Economics, College of Business Administration,
Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[b]Department of Business Administration, College of Business Administration,
Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia
[c]Department of Science and Technology, Mathematics Program University College,
Rania Taif University, Taif, Saudi Arabia
[d]Department of Mathematics, College of Science, Taif University, Taif, Saudi Arabia

Original scientific paper
https://doi.org/10.2298/TSCI22S1437A

This paper aimed to compare the multi-layer perceptron as an artificial neural network and the decision tree model for predicting OPEC crude oil production. Machine learning is about designing algorithms that automatically extract valuable information from data, and it has seen many success stories. The accuracy of these two models was assessed using symmetric mean absolute percentage errors, mean absolute scaled errors, and mean absolute percentage errors. Achieved were the OPEC crude oil production's maximum projected figures. The OPEC crude oil output was also represented by certain descriptive scales and graphs; A comparison was made between the results and the earlier results acquired by the others after the study of the association between the variables revealed statistical significance.

Keywords*: machine learning, artificial neural network, symmetric mean absolute percentage errors, mean absolute percentage error*

Introduction

A recent study uses the ANN approach to create prediction models for dependent variables, to enhance the prediction model for OPEC crude oil production use of ANN techniques was investigated. The prediction model based on an ANN reduces operational expenses and experiment time in research of the analysis of the adsorption behavior of nonionic and anionic individual surfactants. In comparison to the binary model, it also provides accurate forecasts.

The numerous models that were employed in earlier works on earnings prediction are summarized in this section. [1] reviews the significant interest in firm valuation and fundamental analysis that has been shown by earlier scholars in his capital markets study. He

*Corresponding author, e mail: haalrahim@pnu.edu.sa

concluded that these kinds of accounting tests would be helpful in comprehending corporate finance and capital market investment decisions. Therefore, numerous studies that concentrate on conducting empirical experiments to forecast earnings from fundamental data have been published. One of the most popularly used techniques for predicting oil field production [2] is the Arps decline model. The Arps decline model's predictions, meanwhile, are not perfect. In [3], this research used a long-short-term memory (LSTM) network and a gate recurrent unit (GRU) to forecast oil output. Using several data variables, these models forecast oil well production. The created method considers time series and can deal with nonlinear problems. Based on a case study of the two numerical models applied to data acquired from actual oilfields in China and India, the LSTM and GRU were utilized to estimate oil production. The results show that, given different input parameters, LSTM and GRU both have advantages that make them effective approaches for dynamic prediction of oil well production. These strategies are useful and have the potential to be used as a quick and real-time auxiliary basis for planning oil well production [4]. In this paper, the goal of identifying the most advantageous lag in the crude oil price data is given to an artificial neural network model. The forecast is correct up until a significant and abrupt change in the actual data, at which point it becomes difficult to anticipate the precise new price associated with the change. However, the suggested model has effectively taken into account these trends. The outcome is displayed in the figures. This research [5] concentrated on predicting Saybolt color, which is critical for gauging the quality of petroleum products and determining the next step in the process. As inputs, density, kinematic viscosity at 20 °C, sulfur concentration, cetane index, and total acid count were utilized to develop an ANN model with good accuracy in predicting Saybolt color.

Artificial neural network methods

Multilayer perceptron

A multilayer perceptron (MLP), a type of ANN, produces a set of outputs based on a set of inputs. The MLP approach connects a set of dependent variables to a set of predictor attributes, producing prediction models for those variables [6, 7]. The goal of this study's authors [8] was to accelerate the estimation of oil and gas production. They did this by using ML and DL approaches. Several transform functions were added to the models in order to transform the data. The following are the main findings of our investigation [9]. The authors obtained the highest results using ANN, XGBoost, and RNN, with mean R^2 values for oil, gas, and water of 0.9627, 0.9012, and 0.926, respectively. The authors discovered that, whereas the other algorithms performed better with the bespoke dataset, ML methods fared best using the default dataset. Some approaches, such as SVR with a mean R^2 of 0.9014, produced more significant results if the data were standardized prior to the experiment. However, some methods, such as RFR with a mean R^2 of 0.8848, did better with a pure dataset. Normalizing the dataset did not yield good results for either the default or custom datasets; instead, pure and standardized data performed better.

After experimenting with the dataset and examining the findings for each selected method, it is impossible to say that these are the authors' best results. There is still a great deal of room for development to achieve even better outcomes by experimenting with other strategies or a combination of techniques. Non-etheless, given the issue's difficulty, the results we obtained are satisfactory. A specific method known as continuous backpropagation is frequently used to train layered networks known as MLP [10]. The input (the first) layer, the hidden (concealed) layer, and the layer of output are the three node layers that comprise an

MLP network [8]. To evenly distribute the input patterns, a synthetic neural network technique involves changing the weights for a particular training set [11]. The weights are determined throughout the network's learning phase. Three stages are necessary for network training during the learning phase: feedforward of the input training, error computation, and weight calculation. Following training, the network may produce results quickly.

Decision tree

A decision support tool called a decision tree employs a model that resembles decisions and their anticipated results [12] such as event outcomes, resource costs, and tool costs. An algorithm that only uses conditional controls can be shown in one method like this. The use of decision trees in machine learning is frequent. A *test* on an attribute is represented by each internal node, and each branch shows the outcome of the test [13, 14]. A class label sheet is represented by each node (the decision taken after calculating all attributes). The categorization rules are represented by the pathways from the root to the sheet.

The CHAID and exhaustive CHAID algorithms

Exhaustive CHAID was developed by Ritschard [15]. The CHAID algorithm is a modified version of the CHAID decision-tree method. [16] to make up for some of the shortcomings of the latter. Exhaustive CHAID differs from conventional CHAID in that it considers all potential splits on each node and continues splitting even after the best split has been found. There are only two subcategories left after merging all of the predictor variable's categories. The three main steps are merging, stopping, and dividing [17] to make up for some of the shortcomings of the latter. Exhaustive CHAID differs from conventional CHAID in that it considers all potential splits on each node and continues splitting even after the best split has been found [18]. There are only two subcategories left after merging all of the predictor variable's categories. Merging, splitting, and stopping are its three essential processes. These processes are repeated on each node in a decision tree, starting with the root node [17].

The CHAID algorithm

The next algorithm only takes categorical predictors that are nominal or ordinal [19]. Prior to applying the following procedure, continuous predictors are converted into ordinal predictors [20].

For a given set of breakpoints a_1, a_2,…,$a_{k\text{-}1}$. A given x map it to category $C(x)$ as follows (in ascending order):

$$c(x)=\begin{cases}1,\ x\le a_1\\ k+1,\ a_k<x<a_k+1,\ k=1,...,k-2\\ k,\ a_{k-1}\le x\end{cases}$$

desired number of bins, the breakpoint point is computed. Do a rank calculation for x. The ranks are computed while factoring in frequency weights. The average rank is used if there are ties.

In ascending order, write the ranking and corresponding values:

$$\{r_i,x_i\}_{i=1}^{n}$$

for k = 0 to (k – 1), set

$$I_K=\left\{i:L_{r(i)}\frac{k}{N_{f+1}}\ j=k\right\}$$

where [x] denotes the lower integer of x.

If I_K is not empty $i_k = \max\{i: i \in I_k\}$

The breakpoints are configured to correspond to the x values of the, except the greatest.

The next algorithm only takes categorical predictors that are nominal or ordinal. Prior to applying the following procedure, continuous predictors are converted into ordinal predictors.

A given x mapped into category $C(x)$ in the following manner for a given set of breakpoints a_1, a_2,.., a_{k-1}, (in increasing order).

The breakpoint is calculated as follows if k is the desired number of bins.

Do a rank calculation for x. The ranks are calculated using frequency weights. If there are ties, the average rank is used, calculated using frequency weights. If there are ties, the average rank is used. Write the rank and corresponding values in ascending order, as the floor integer of x is denoted by the term set for $k = 0$ to $(k - 1)$. should not be empty. The breakpoints are configured to correspond to the respective x values.

The accuracy measurement

To assess the forecast's accuracy, we employed the symmetric mean absolute percentage error (sMAPE), mean absolute scaled error (MASE), and mean absolute percentage error (MAPE). The formulas below can be used to determine sMAPE, MASE, and MAPE [21, 22]:

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{|y_i|+|\hat{y}_i|} \tag{1}$$

$$MASE = \frac{\frac{1}{n}\sum_{i=1}^{n}|e_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|y_i - y_{i-1}|} \tag{2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \tag{3}$$

Numerical results

To classify and predict the OPEC crude oil production based on historical from OPEC annual report from 1970 to 2017 data [23] we create a decision tree and multi-layer perceptron.

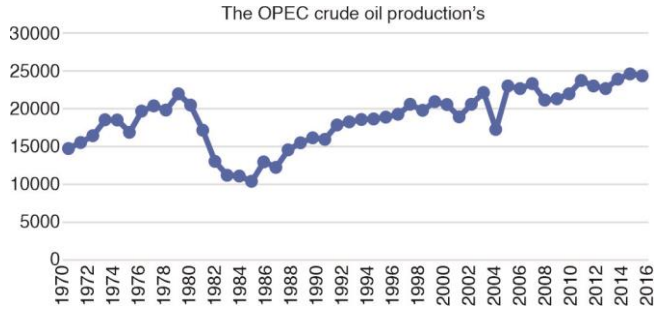

Figure 1. The graph depicts the time series of OPEC crude oil production

Decision tree

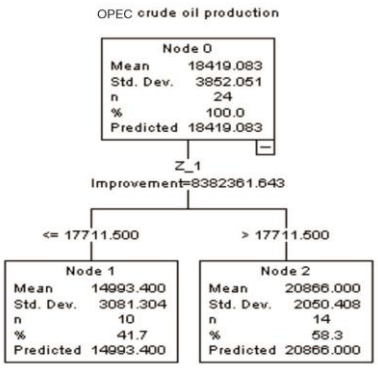

Figure 2. Structure of a training sample decision tree with two nodes

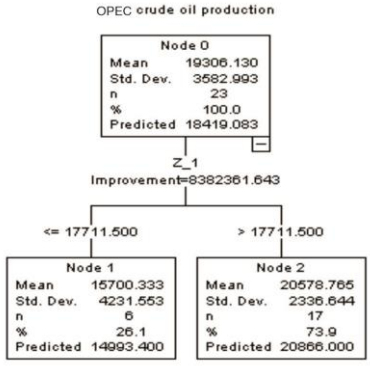

Figure 3. Structure of a testing sample decision tree with two nodes

Multi-layer perceptron

Table 1. Case processing summary

Case processing summary			
		N	%
Sample	Training	24	51.1
	Testing	23	48.9
Valid		47	100.0
Excluded		2	
Total		49	

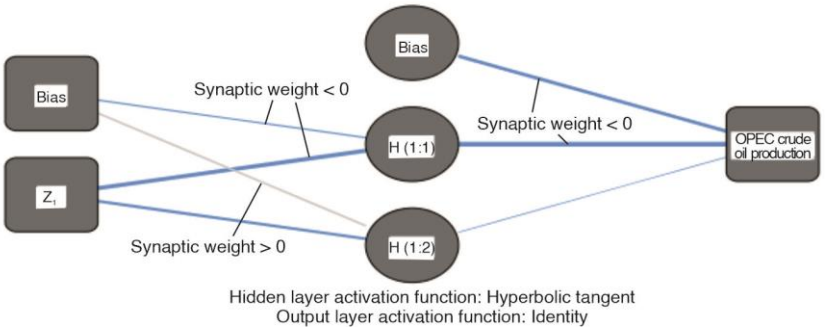

Figure 4. Shows the hidden layer activation function is hyperbolic tangent and the output layer activation function is identity

The accuracy measurements

Table 2. Show the accuracy measurements

	Multi-layer perceptron	Decision tree
SMAPE	0.007118118	0.021889569
MSE	0.031140986	0.111897868
MAPE	0.00753731	0.026425463

Conclusions

In this study, decision trees and ANN models are utilized to predict OPEC crude oil production using a machine learning approach. Divide the data by 50.1% for training and 49.9% for testing in the decision tree model as the best option, see tab. 1. The decision tree structure, which has two nodes for the testing sample and the training sample, is shown in figs. 1 and 2. Node one had a mean of 1499.4, and node two had a mean of 2050.408 for the training sample. Node one had a mean of 15700.333, and node two had a mean of 20578.765 for the testing sample. In fig. 3, there are no numbers or weights associated with the lines linking the nodes. Rather, the color and width of these lines instead represent their weights; thick lines indicate a weight that considerably deviates from zero and are colored blue when they do so and grey when they do not. The output layer's two nodes, which stand for completion and non-completion, contain bias boxes that are designed to correct systematic prediction errors. The activation function is hyperbolic tangent and the output layer activation function is identity in fig. 4. Table 2 shows that the outcomes of the simulation study demonstrate the efficacy of the ANN model's data selection strategy.

Acknowledgment

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University, through the Research Funding Program, Grant No. (FRP-1443-20)

References

[1] Kothari, S., Capital Markets Research in Accounting, *Journal of Accounting and Economics, 31* (2001), 1-3, pp. 105-231
[2] Cheng, Y., Yang, Y., Prediction of Oil Well Production Based on the Time Series Model of Optimized Recursive Neural Network, *Petroleum Science and Technology, 39* (2021), 9-10, pp. 303-312
[3] Pavlyshenko, B. M., Machine-Learning Models for Sales Time Series Forecasting, *Data, 4* (2019), 1, p. 15
[4] Gupta, N., Nigam, S., Crude Oil Price Prediction Using Artificial Neural Network, *Procedia Computer Science, 170* (2020), pp. 642-647
[5] Salehuddin, N. F., *et al*., A Neural Network-Based Model for Predicting Saybolt Color of Petroleum Products, *Sensors, 22* (2022), 7, 2796
[6] Elhag, A. A., Abu-Zinadah, H., Forecasting Under Applying Machine Learning and Statistical Models. *Thermal Science, 24* (2020), Suppl. 1, pp. S131-S137
[7] Jarrah, M., Salim, N., A Recurrent Neural Network and a Discrete Wavelet Transform to Predict the Saudi Stock Price Trends, *International Journal of Advanced Computer Science and Applications, 10* (2019), 4
[8] Chattopadhyay, G., *et al*., MLP Based Predictive Model for Surface Ozone Concentration Over an Urban Area in the Gangetic West Bengal During Pre-Monsoon Season, *Journal of Atmospheric and Solar-Terrestrial Physics, 184* (2019), Mar., pp. 57-62
[9] Romero, E., Sopena, J. M., Performing Feature Selection with Multilayer Perceptrons, *IEEE Transactions on Neural Networks, 19* (2008), 3, pp. 431-441
[10] Orhan, U., *et al*., EEG Signals Classification Using the K-Means Clustering and a Multilayer Perceptron Neural Network Model, *Expert Systems with Applications, 38* (2011), 10, pp. 13475-13481
[11] Hamdi, M., *et al*., Forecasting and Classification of New Cases of COVID 19 Before Vaccination Using Decision Trees and Gaussian Mixture Model, *Alexandria Eng. Journal, 62* (2023), Jan., pp. 327-333
[12] Borrego-Morell, J. A., *et al*., On the Effect of COVID-19 Pandemic in the Excess of Human Mortality, The case of Brazil and Spain, *PloS One, 16* (2021), 9, e0255909
[13] Abo-Dahab, S., *et al*., Free Convection Effect on Oscillatory Flow Using Artificial Neural Networks and Statistical Techniques, *Alexandria Engineering Journal, 59* (2020), 5, pp. 3599-3608
[14] Ranka, S., Singh, V., CLOUDS: A decision tree classifier for large datasets, *Proceedings*, 4th Knowledge Discovery and Data Mining Conference, New York, USA, 1998

[15] Ritschard, G., CHAID and Earlier Supervised Tree Methods, in: *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences*, Routledge, Oxford, UK, 2013, pp. 70-96
[16] Biggs, D., *et al*., A Method of Choosing Multiway Partitions for Classification and Decision Trees, *Journal of Applied Statistics, 18* (1991), 1, pp. 49-62
[17] Novita, R., *et al*., Identifying Factors that Influence Student Failure Rate Using Exhaustive CHAID (Chi-Square Automatic Interaction Detection), *Proceedings*, 3rd International Conference on Information and Communication Technology (ICoICT), Melaka Campus, Kuala Lumpur, Malaysia, 2015
[18] Sugumaran, V., *et al*., Feature Selection Using Decision Tree and Classification Through Proximal Support Vector Machine for Fault Diagnostics of Roller Bearing, *Mechanical Systems and Signal Processing, 21* (2007), 2, pp. 930-942
[19] Diaz-Perez, F. M., Bethencourt-Cejas, M., CHAID Algorithm as an Appropriate Analytical Method for Tourism Market Segmentation, *Jou. of Destination Marketing & Management, 5* (2016), 3, pp. 275-282
[20] Trujillano Cabello, J., *et al*., Stratification of the Severity of Critically Ill Patients with Classification Trees, *BMC Medical Research Methodology, 9* (2009), 83, pp. 1-12
[21] Vijay, G., *et al*., Performance and Emission Prediction in a Biodiesel Engine Fuelled with Honge Methyl Ester Using RBF Neural Networks, *International Journal of Mechanical and Mechatronics Engineering, 9* (2015), 6, pp. 976-981
[22] Makridakis, S., *et al*., Statistical and Machine Learning forecasting Methods: Concerns and Ways Forward, *PloS One, 13* (2018), 3, e0194889
[23] Lyons, C., The Organization of the Petroleum Exporting Countries (OPEC) (www. opec. org), *Journal of Business & Finance Librarianship, 14* (2009), 2, pp. 181-187

Paper submitted: October 15, 2022
Paper revised: November 22, 2022
Paper accepted: November 28, 2022