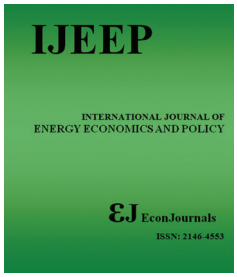


International Journal of Energy Economics and Policy

ISSN: 2146-4553

available at http: www.econjournals.com

International Journal of Energy Economics and Policy, 2024, 14(2), 594-602.

Modelling Stock Prices of Energy Sector using Supervised Machine Learning Techniques

Mimoun Benali, Lahboub Karima*

Laboratory of Research and Studies in Management, Entrepreneurship and Finance (LAREMEF), National School of Commerce and Management of Fez, Sidi Mohamed Ben Abdellah University, Fes 30050, Morocco. *Email: lahboubkarima.encg@gmail.com

Received: 26 October 2023 **Accepted:** 15 February 2024 **DOI:** https://doi.org/10.32479/ijeep.15553

ABSTRACT

This paper aims at comparing the performance of the different state-of-the-art machine learning techniques in anticipating the performance of stock prices of the energy sector. The data collected cover the period from January 2020 to February 2023 with a daily frequency for the three most imported refined petroleum products in Morocco and trained four regression machines learning (linear regression, lasso regression, ridge regression, and SVR) and four classifiers machine learning (logistic regression, decision tree, extra tree and Random Forest) so that anticipating 1 day ahead prices direction can take place no matter whether they are negative or positive prices. The performance of regression algorithm is then evaluated using different evaluation metrics, especially MSE, RMSE, MAE, MAPE and R2 to evaluate the performance of regression algorithm while precision, recall and F1 scores are used to evaluate the performance of classifiers algorithm. The outcomes propose that the performance of linear regression and ridge regression takes place equally and outperform other single regression that is lasso regression and SVR for-1-day predictions as a whole. In addition to that, we have come to find that in the classifiers, algorithms group all machine learning display similar predictive accuracy, this is on one hand. On the other hand, the best of them is the logistic regression. In brief, this study suggests that all performance metrics are significantly improved by ensemble learning. Therefore, this study proves that critical information affecting stock movement can be captured by utilizing historical transactions.

Keywords: Machine Learning, Stock Price, Energy Sector, Regression, Price Prediction
JEL Classifications: Q4, C54

1. INTRODUCTION

One of the key factors for the development of economy and social progress of all countries around the world is the Energy. The latter's importance has grown to the extent that it has become one of the major concerns for the international community in the present days. The global energy demand, especially in some developing countries such as Morocco. which has few energy resources as well as the rely holistically on imports for its supply of modern energy resources to meet the increasing needs associated with its economic and demographic growth, is rapidly growing. Energy goods, particularly natural gas and oil, are increasingly subject to disruption following either natural disasters, armed conflicts or nationalist policies practiced by the main producing countries. Thus, by the end on 2022, energy prices in Morocco soared to record levels, which was due to suffering from a difficult external environment, especially from the effect of the war between Russia and Ukraine. According to the Ministry of Energy Transition and Sustainable Development, the imports of energy recorded a strengthening in value of around 116.3% (or about + 69 billion DHs) to be about 128.3 billion DHs at the end of the period January-October 2022 instead of about 59.3 billion DHs at the end of October 2021. A rise in purchases in value of all product groups is reflected due to this trend. At the end of October 2022, the level of the gross bill of energy remains decreased above the gross energy recorded during the same period of the last years before the corona virus crisis: from 2017 to 2019 as an example. Additionally, since Samir stopped working, Morocco no longer

imports crude oil and the imported refined petroleum products come mainly from Russia, Italy, France, Spain and Netherlands. Furthermore, three companies from 20 distributors, which are Afriquia and the subsidiary of the Akwa group (Vivo Energy or Shell) and Total Morocco, account for more than 70% of the overall imports. To conclude, Global economic developments, geopolitical and environmental factors have led to a significant increase in the volatility of energy prices. Consequently, the ability to predict energy prices becomes very important.

Forecasting the stock price is of great importance for not only energy policy-makers, market participants like large corporate producers, manufacturers, large energy consuming firms as it can seriously affect its competitiveness, viability and future profitability but also portfolio diversification and energy risk management. Many attempts have been made to show the behavior of oil prices. While estimating the price of oil, two key categories, which are conventional statistics and econometric approaches, are taken into account. However, the classical economic models manage to cope with a large number of variables in big data sets; the limit order book and news stories as an example. Thus, stock prices are nonstationary, noisy and exhibit non-linear dynamics that cannot be captured by simple linear models (Abu-Mostafa and Atiya, 1996), and that is exactly what influence the model's performance. Having a solution or set of solutions for most problems in economic domain entails the ability of applying machine learning approaches. Techniques of Machine learning (ML) possess these capabilities as it was explored in Varian (2014). Machine learning models also have better prediction performance robustness in comparison with to traditional methodologies of statistics and econometric models (Patel et al., 2015; Bustos and Pomares-Quimbaya, 2020; Shobana and Umamaheswari, 2021). Rather than being explicitly programmed, especially in oil price modelling, machine learning refers to a collection of algorithms that enable systems to learn from experience and data (Cohen, 2020; Tang et al., 2020).

In a range of areas, the burst of applications of ML is due to the accessibility to plenty of data, the improvement in algorithms, and the increase in computer power. Finding the different models that effectively go hand in hand with the data points and predicting the stock price using the four regression machines learning (linear regression, lasso regression, ridge regression and SVR) as well as the four classifiers machine learning (logistic regression, decision tree, extra tree and random forest) to know which gives the best in terms of accuracy and performance is what this project aims at. Similarly, comprehensive and accurate national energy planning must take place for maintaining sustainable development in Morocco. Consequently, to make sure that Morocco is on track with its 'energy efficiency' strategy by 2030, forecasting stock prices is needed. In the energy sector, and our knowledge, this article remains among the first studies of oil price forecast in Morocco. For this reason, and as regards prediction accuracy, time saving and robustness, and doing a comparative analysis on the performance of the model with other prediction models have been presented somewhere else by some researchers, as well as the proposed model on python programming environment and evaluating the performances of the model, this study aims at formulating a new more intelligent predictive model using different kinds of machine learning and enhancing three performance of oil price prediction in Morocco.

What we are going to see through this paper is as follows: chapter one presents some reading in the previous studies, chapter two describes the area of study, data collection procedure, and partitioning process. Chapter three is about research methodology; it is basically about upon which techniques this research takes place. Then chapter four which provides the mains results.

2. LITERATURE REVIEW

The intrinsic complexity of oil market mechanisms always makes oil price prediction an intractable task, and the debate on understanding the behavior underlying oil prices has been renewed by the recent oil shocks and their far-reaching consequences. Generally speaking, the categorization of the previously done research on oil price forecasting can be divided into three main categories: econometric models, machine learning or deep learning methods and hybrid approaches. Because of the fact that the oil price series are usually nonlinear and non-stationary time series, which is interactively affected by many factors, Zhang et al. (2008) revealed three main factors (short-term, medium-term, and long-term) which affect the price of the crude oil volatility. It is typically driven by supply from exporting countries and demand from industrialized countries. Moreover, the oil price is influenced by other factors like GDP, exchange rate and financial asset as well as by economic crises and political events Hamilton (2003); Hooker (1996); Narayan et al. (2014); Narayan and Sharma (2014); Pradhan et al. (2015); Phan et al. (2015b); Narayan and Narayan (2014); Dowling et al. (2014).

The previously-mentioned variables affecting crude oil price have the characteristics of complex nonlinearity, dynamic volatility and high irregularity Watkins and Plourde (1994). remain the main reason behind predicting oil price accurately is rather challenging. Detecting nonlinear chaotic dynamics in time series data was dependent on nonlinear time series analysis on which the early studies were based. In the recent past decades, traditional statistical and econometric models such as linear regression, co-integration analysis, error correction models, GARCH models, naïve random walk, probabilistic model and vector auto-regression have also been widely applied to oil price forecasting Hooman (2020). The models we have just seen usually provides good prediction results when the price series under study and analysis is linear or partially linear. As usual, and due to the fact that oil price is formulated by complex factors that have different interactive effects between themselves as we mentioned, there is always a great deal of nonlinearity and irregularity.

The use of the traditional statistical and economic models remains the reason behind having poor prediction performance as lot of studies and experiments have recently shown (Weigend and Gershenfeld, 1994). Being built on linear assumptions and being unable to capture the nonlinear patterns hidden in the oil price series, the traditional statistical and economic models remain the reasons behind this phenomenon. Some linear and emerging artificial

intelligent models, nonlinear regression, artificial neural networks, deep learning and genetic expression programming as examples, provide powerful solutions to nonlinear oil price prediction simply because the limitations of the traditional statistical and econometric models (Wang and Tao, 2008; Zhang et al., 2017; Wu et al., 2019; Wang et al., 2021; Butler et al., 2021; Fan et al., 2021; Li et al., 2019; Zhao et al., 2017). Using ML tools in the crude oil price forecasting process, researchers generally share the same point of view which is that advanced and hybrid ML tools are superior to traditional statistical tools. Hybrid methods integrate the methods mentioned above and thus utilize their advantages synthetically (Hooman, 2020; Hooman and Seyed, 2020). To enhance the forecasting performance, Zhou et al., (2019) accumulate time series disintegrating models and AI models.

There are two reasons behind why hybrid methods usually achieve better results than single models performance and they are as follow: in the price time series, we have time-series approaches or econometric models specialize in capturing the linearity and volatility, on the one hand while, on the other hand, the specialization in nonlinear and non-stationary characteristics goes to AI models. The division of utilization of news to enhance crude oil price forecasting, according to the recent research, can be into two: the first one is that news sentiment (Li et al., 2017; Zhao et al., 2019) or Google trends Wu et al. (2021) are utilized mainly as auxiliary features while the second one resides in the concentration on the news contexts. For constructing predictive factors, Bai et al., (2021) use the Latent Dirichlet Allocation topic model while for extracting text features from online news headlines automatically, Binrong et al. (2021) use Convolutional Neural Network. By comparing different ARIMA and NN architectures, or through the application of different decomposition and/or joint techniques, recent studies have paid more attention on the determination of the optimal forecasting technique (Tseng et al., 2002; Shambora and Rossiter, 2007; Kadkhodaie-Ilkhchi et al., 2009; Christodoulos et al., 2010; Jammazi and Aloui, 2012; Paz-Marín et al., 2012).Technologies with increasing number of real-world applications including finance and economics are being evolved by neural networks and genetic algorithms (Lisboa and Vellido, 2000; Chen, 2002; Sermpinis et al., 2012), and it has been successfully applied problems in biology mining and computing (Dehuri and Cho, 2008; Lopez and Weinert, 2004a; 2004b; Margny and El-Semman, 2005). Genetic expression programming is a new technique and its application are quite limited in the energy sector as well.

3. DATA

In this empirical paper which inspects the predictive power of the different models of the machine learning applied to the Moroccan energy sector. The three selected representative stocks as analysis subjects from Casablanca Stock Exchange/Amsterdam Stock Exchange are: Afriquia, the subsidiary of the Akwa group, Total Morocco and Vivo Energy or what is called Shell. Opening Price, Closing Price, Highest Price and Lowest Price are what the stock historical transaction data set includes. The period from January 2020 to February 2023 is the period when data collection takes place with a daily frequency and the prediction of the next day opening price is our target. In this respect, target = current day closing price – next day opening price. The aim is testing the performance of regression and classifier machine learning algorithms varies as a function of the prediction horizon. Formally, we generate the output variable for each stock as follow:

$$Target_t = \begin{cases} \Delta\ 1;\ if\ P_{(t+n)} > P_t \\ \Delta\ 0;\ otherwise \end{cases}$$

The output variable is either equal to one (i.e. class 1) if the price at t+n is positive or zero otherwise (i.e. class 0). where Pt is the adjusted closing price at time t for daily prediction horizon.

With the additional support package scikit-learn (version 0.24.1) to compute and run the ML algorithms, python (version 3.8.3) is the programming environment used in this study.

- Training: Set which allows the algorithm to perform the task to predict or classification by learning and improving from the training set. The latter represents 88% of the observations;
- Validation: Set allows the model parameters to be optimized and it represents 10% of the data.
- Test: Set, which is mainly used to evaluate the predictive performance, covers the remaining 2% of the observations.

4. METHODOLOGY

4.1. Machine Learning Models

The use of a machine learning model is what helps to find patterns and decisions from the previously unseen dataset. Generally speaking, supervised learning, unsupervised learning and reinforcement learning are the classifications of most machine learning techniques. supervised learning, where the chosen algorithm tries to fit the target using the given input, is the used type of machine learning in this study. In other words, to find the relationship which has the best predictive power, the supervised learning algorithms are given the historical data. There are two varieties of supervised learning algorithms which are regression and classification algorithms. Regression-based supervised learning methods mission is to predict outputs based on input variables while classification-based supervised learning methods identify which category of set of data items belong to. Additionally, classification algorithms are probability-based. This means that the outcome is the category for which the algorithm finds the highest probability that the dataset belongs to; whereas, regression algorithms estimate the outcome of problems that have an infinite number of solutions which is referred to as the continuous set of possible outcomes.

4.1.1. Machine learning regression

4.1.1.1. Linear Regression

Multiple linear regression is a known of its use as a statistical analysis method to estimate the marginal effects of carefully chosen independent variables on explanatory dependent variables. The ordinary least squares method is a simple method to estimate the relationship between the independent variable and the explanatory variable in multiple linear regression.

4.1.1.2. Lasso Regression

The lasso model, which stands for Least Absolute Shrinkage where the data values are shrunk towards a central point as the mean and the Selection Operator, is another modification of the least square method. It is used to estimate the sparse coefficients like models with few parameters. To automate certain parts of model selection, like variable selection/parameter estimation, and for the models which show high levels of multi-collinearity. This particular type of regression is well-suited. L1 regularization technique is used by Lasso Regression which, it is used when we have more features because it automatically performs feature selection. Rather than the sum of the squared coefficients, it adds the total of the absolute values of the coefficients.

4.1.1.3. Ridge Regression

Ridge regression is a regularization approach, which belongs to the class regression tool that uses L2regulation, is the way to combat the issue of overfitting in linear regression models. By adding a penalty term to the coast function of linear repression, the size of the coefficient is reduced and the overfitting is prevented. The penalty term regulates the magnitude of the coefficients in the model and is proportional to the sum of squared coefficients. The penalty term's value is raised when the coefficients shrink towards zero lowing the model's variance.

4.1.1.4. Support vector regression

Support vector regression, which is used for regression tasks, is a type of support vector machine. Finding a function that approximates the relationship between the input variables and the continuous target variable is the goal of the SVR while minimizing the prediction error. To define how much error is acceptable in our model and to find an appropriate line to fit the data, SVR provides us this flexibility. This takes place by mapping the input variables to a high-dimensional feature space and finding the hyperplane which maximizes the margin or the distance between the hyperplane and the closest points as well minimizing the prediction error.

4.1.2. Machine learning classification

4.1.2.1. Logistic Regression

Logistic regression is approximately the same as the Linear Regression except that how they are used. Linear regression is usually used to solve Regression problems while Logistic regression is used to solve the classification problems. It is called regression simply because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class, which is a mathematical function used to map the predicted values to probabilities and S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never the same at those limits.

4.1.2.2. Decision Tree

Based on a given conditions, it remains graphical representation for getting all the possible solutions to a given problem. The reason behind being called a decision tree is that it is the same as a tree and it starts with the root nodes which are the decision node and the leaf node. To make any decision, decision nodes, which have different branches, are used; whereas, leaf nodes are the outputs of those decisions and have no branches. Before following the branches and jumping to the next node based on comparison, this algorithm compares the values of root attribute with record attribute first, and the algorithm, again, compares the attribute value with the other sub-nodes and move further, for the next node. This procedure continues until it reaches the leaf node of the tree.

4.1.2.3. Extra Tree

The extra trees are similar to random forest but can be faster. Even though the sampling for each tree is random and without replacement, the random forests algorithm creates many decision trees. There are also selected randomly specific numbers of features from the total set of features, and the most important and unique characteristic of extra trees is the random selection of a splitting value for a feature. What makes the tree diversified and uncorrelated is the random selection of a split values the algorithm does instead of calculating a local optimal value using Gini or entropy.

4.1.2.4. Random Forest

A number of decision trees on different subsets of the given dataset as well as taking the average to improve the predictive accuracy of the dataset is contained by Random Forest; the classifier type of machine learning. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The bigger the number of the trees in the forest, which leads to preventing the problem of overfitting, the higher accuracy becomes.

4.2. Evaluation Metrics

This paper aims at giving the comparison results of the above ML models in terms of confusion matrix and other performance measure for the sake of measuring the prediction performance of them (ML models).

4.2.1. Confusion matrix

A matrix, which is with 4 different combinations of predicted and actual values, describing the performance of a classification model on a set of test data depicts the confusion matrix.

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{Recall}=\frac{TP}{TP+FN}$$

$$F1=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}=\frac{2*TP}{2*TP+FP+FN}$$

When we talk about true positive (TP): Observation is predicted positive and is actually positive in the same way, false positive (FP): Observation is predicted positive and is actually negative. While, true negative (TN): Observation is predicted negative and is actually negative. Observation is predicted negative and is actually positive when it comes to False negative (FN).

4.2.2. Performance measure

In the sake of evaluating prediction on regression machine, there are actually several evaluation metrics that exist for the matter. Coefficient of determination 'R^2', mean square error 'MSE', and root mean square error 'RMSE', mean absolute error 'MAE' mean absolute percentage error 'MAPE'. Strengths and weaknesses of each method remains the factor behind having different prediction metrics. What makes RMSE and MAE scale-dependent error measures is that they do not allow comparison between point predictions across different scales, and the use of the percentage-based error measure MAPE makes the probability of having errors smaller when the predictor variable is lower. The following are the equations for this metrics:

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i|}{N}$$

$$MSE = \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}$$

$$MAPE = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$

Where N denotes the number of samples, $\hat{y}_i$ is the predicted value of the model, y_i m eans the true value of the response, and $\bar{y}_i$ represents the average estimate.

5. EMPIRICAL RESULTS

5.1. Evaluation of Regression Algorithms

To find out which regression algorithm suits these kinds of datasets so that researchers find it helpful for their future studies, four different algorithms on the data of three different companies applied to the Moroccan energy sector are compared in this study. Table 1 and the Figures 1-4 represents results of these algorithms of the stock as a whole, and each column represents a stock and each row represents machine learning algorithms.

On one hand, when we applied the mentioned algorithms on the data of the Afriquia Company, linear regression, lasso regression and ridge regression show better results on training data and data test than SVR. On the other hand, SVR shows better results on validation data in terms of the MSE, the RMSE, the MAE, the MAPE and the R2 criterions than linear regression, lasso regression and ridge regression.

Table 1: Evaluation of regression algorithms

Model	M	AFR	TTL	SHELL
Linear regression				
Training	MSE	10,496.26	359.26	0.04
	RMSE	102.45	18.95	0.19
	MAE	74.03	13.1	0.14
	MAPE	1.63	0.96	0.73
	R^2	0.97	1	1
Validation	MSE	19,579.25	649.05	0.35
	RMSE	139.93	25.48	0.6
	MAE	106.83	18.45	0.38
	MAPE	2.36	1.85	2.3
	R^2	−0.22	0.98	0.89
Test	MSE	8702.71	649.32	0.02
	RMSE	93.29	25.48	0.14
	MAE	72.31	20.2	0.12
	MAPE	1.64	1.52	0.44
	R^2	−0.59	0.85	0.84
Lasso regression				
Training	MSE	10,512.54	371.95	0.07
	RMSE	102.53	19.29	0.27
	MAE	73.95	13.22	0.22
	MAPE	1.63	0.97	1.17
	R^2	0.97	1	1
Validation	MSE	19,511.46	642.86	0.37
	RMSE	139.68	25.35	0.61
	MAE	106.74	18.69	0.4
	MAPE	2.36	1.88	2.44
	R^2	−0.22	0.98	0.98
Test	MSE	8424.22	696.41	0.04
	RMSE	91.78	26.39	0.19
	MAE	69.95	21.21	0.16
	MAPE	1.59	1.6	0.6
	R^2	−0.54	0.83	0.73
Ridge regression				
Training	MSE	10,496.26	359.26	0.04
	RMSE	102.45	18.95	0.19
	MAE	74.03	13.1	0.14
	MAPE	1.63	0.96	0.73
	R^2	0.97	1	1
Validation	MSE	19,579.52	649.05	0.35
	RMSE	139.93	25.48	0.59
	MAE	106.83	18.45	0.38
	MAPE	2.36	1.85	2.3
	R^2	−0.22	0.98	0.98
Test	MSE	8702.71	649.32	0.02
	RMSE	93.29	25.48	0.14
	MAE	72.31	20.2	0.12
	MAPE	1.64	0.85	0.44
	R^2	−0.59	0.85	0.84
SVR				
Training	MSE	278,691.15	365.38	0.06
	RMSE	527.91	19.11	0.24
	MAE	431.28	13	0.18
	MAPE	9.8	0.95	0.92
	R^2	0.27	1	1
Validation	MSE	14,632.06	365.38	0.37
	RMSE	120.96	19.11	0.61
	MAE	93.98	13	0.4
	MAPE	2.07	1.83	2.45
	R^2	0.09	0.98	0.98
Test	MSE	14,716.95	631.23	0.02
	RMSE	121.31	25.14	0.16
	MAE	110.05	20.52	0.12
	MAPE	2.54	0.85	0.45
	R^2	−1.69	0.85	0.81

Table 2: Evaluation of classification algorithms

Model	M	TTL	AFR	SHELL
Logistic regression				
Training	Precision	0.77	0.52	0.85
	Recall	0.68	0.51	0.85
	F1	0.69	0.43	0.85
Validation	Precision	0.68	0.73	0.81
	Recall	0.68	0.61	0.79
	F1	0.68	0.58	0.79
Test	Precision	0.86	0.83	0.76
	Recall	0.6	0.62	0.71
	F1	0.58	0.6	0.73
Decision tree				
Training	Precision	0.98	0.91	1
	Recall	0.96	0.89	1
	F1	0.97	0.89	1
Validation	Precision	0.6	0.43	0.72
	Recall	0.59	0.43	0.72
	F1	0.57	0.43	0.72
Test	Precision	0.5	0.25	0.58
	Recall	0.5	0.33	0.61
	F1	0.49	0.29	0.51
Extra tree				
Training	Precision	0.98	0.91	1
	Recall	0.96	0.89	1
	F1	0.97	0.89	1
Validation	Precision	0.61	0.41	0.61
	Recall	0.6	0.41	0.61
	F1	0.59	0.41	0.61
Test	Precision	0.5	0.25	0.4
	Recall	0.5	0.33	0.36
	F1	0.49	0.29	0.36
Random forest				
Training	Precision	0.97	0.9	1
	Recall	0.97	0.9	1
	F1	0.97	0.9	1
Validation	Precision	0.61	0.3	0.74
	Recall	0.6	0.29	0.73
	F1	0.59	0.29	0.73
Test	Precision	0.33	0.45	0.76
	Recall	0.5	0.46	0.8
	F1	0.4	0.45	0.77

When it comes to the data of Total Morocco Company, all machine learning algorithms show better results on training data in all MSE, RMSE, MAE, MAPE, and R2 criterions but linear regression, and ridge regression still the best. On the validation of the data, the SVR, again, show the best results. The SVR, show the best results and perform best in the MSE, the RMSE, and the MAE compared by linear regression and ridge regression on data test.

While training the algorithms with the data of Shell Company, linear regression and ridge regression show best results on training data while SVR and lasso regression performed worst. On validation test and data test all machine learning algorithms show best results but linear regression, and ridge regression still the best.

Both of linear regression and ridge regression fit better on training data for all the companies; whereas, the SVR and lasso regression, as examples, do not. Optional results, when the relationship between the independent variables and the dependent variables which are almost linear, are shown by the implementation of linear regression on the base of statistical model. Furthermore, results on data validation is shown better using the SVR. Additionally, linear regression, ridge regression and the SVR fit better on testing data for all companies than lasso regression.

5.2. Evaluation of Classification Algorithms

Precision shows that the number of positive predictions is well made. To put clear, it is the number of the well predicted positives (True Positives) divided by all the predicted positive (True Positive + False Positive). The higher it is, the more the Machine Learning Model minimizes the number of False Positives. When the precision is high means that the majority of the positive predictions of the model are well predicted positives.

The percentage of positives well predicted by our model is given by recall. it is the number of well predicted positives (True Positive) divided by the total number of positives (True Positive + False Negative). The higher it is, the more Learning Machine Model maximizes the number of True Positives. When the recall is high, it rather means that it will not miss any positive**.** Nevertheless, any information about its prediction quality on the negatives is given.

A combination of precision and recall is used by the F1-score metric. Indeed, the F1 score is the harmonic mean of the two. A high F1 score symbolizes a high precision and a high recall. This presents a good balance between precision and recall, and it gives good results on problems. A low F1 score does not tell something except that whatever about performance at a threshold. Low recall means we did not try to do well on most of the set test as a whole, and among the cases we identified as positive cases, low precision means that we did not get many of them right, but low F1 does not tell which case. On the first hand, having high F1 means we likely have high precision and recall on a large part of the decision. On the other hand, having low F1 means that neither the problem is clear (low precision or low recall), nor the model suffers from type-I or type-II error.

Table 3 show that Precision, recall, and F1 show higher values on training and validation test when they these algorithms are applied on the data of Total Morocco Company. These values are towards 1 signify that all models did not miss any true positives.

A law score (+<0.5) of precision, recall and F1 on decision tree, extra tree and random forest is shown on testing test. At the same time, this means that this classifier models has a high number of false positives. This is may be due to a result of imbalanced class and untuned model hyper parameters. Table 2 suggests that logistic regression outperforms all other single classifiers (i.e decision tree, extra tree and random forest) in terms of precision, recall and the F-1 for 1-day-ahead forecasting.

Precision, recall and the F1 show higher values on training data when applying these algorithms on the data of Afriquia Company. These values its towards 1 signify that decision tree, extra tree and random forest models do not miss any true positives. Precision, recall and F1 show a low score (<0.5) on validation and testing data, which means that this classifier models has a high number of false positives. This can be the result of an outcome of imbalanced class or untuned model hyper parameters. Table 2 suggests that, in terms of precision, recall and the F1 for 1-day-ahead forecasting

Figure 1: Future day prediction - Linear regression

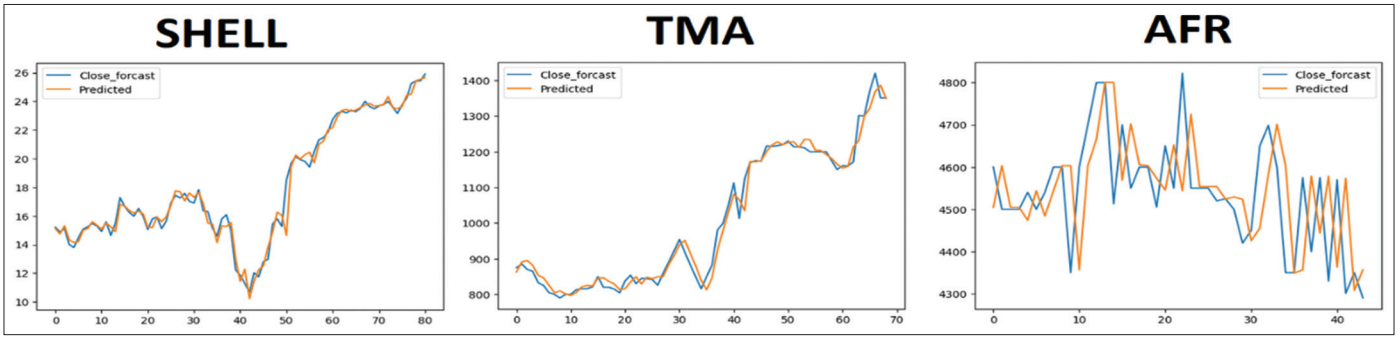


Figure 2: Future day prediction – Lasso regression

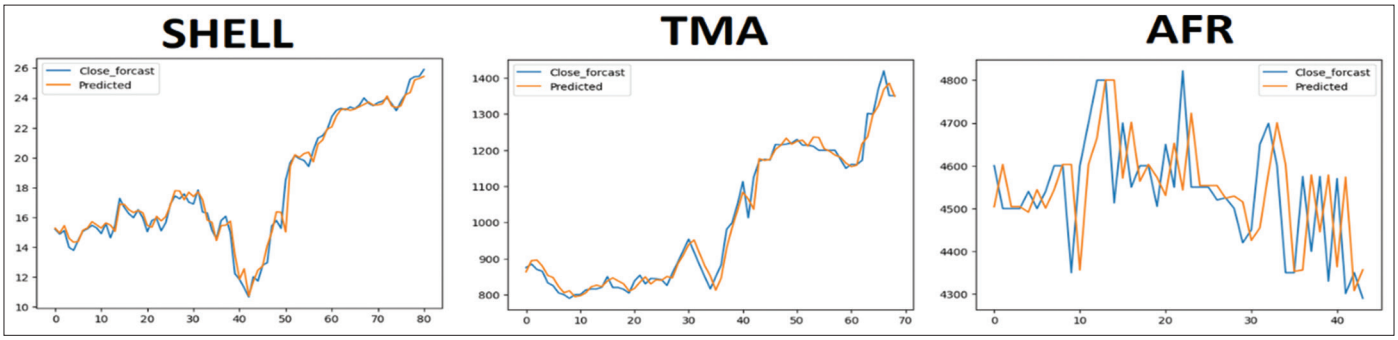


Figure 3: Future day prediction – Ridge regression

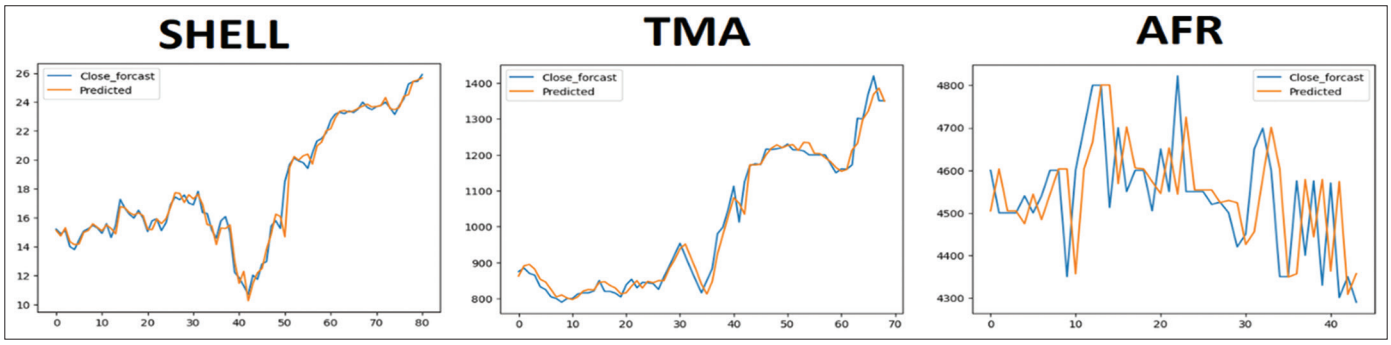
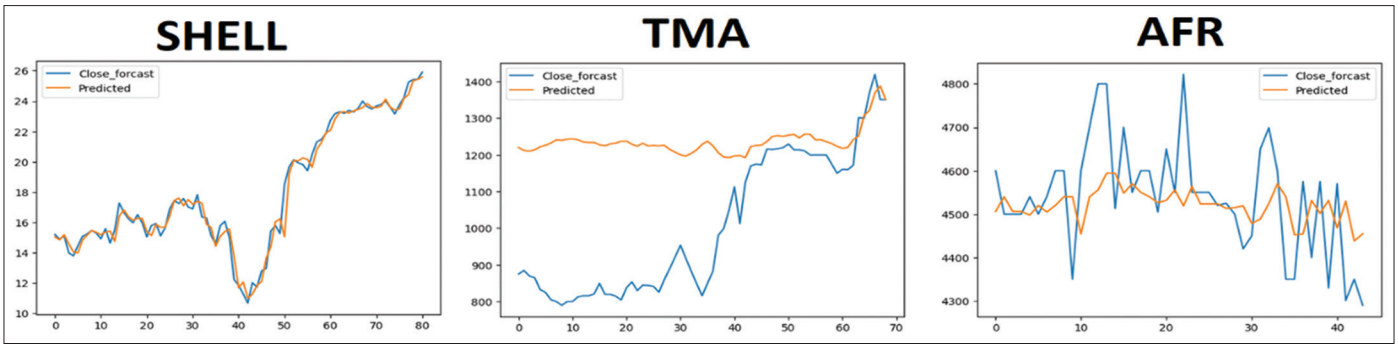


Figure 4: Future day prediction – SVR



on training, validation and testing data, logistic regression outperforms all other single classifiers models (i.e decision tree, extra tree and random forest).

Lastly, logistic regression, decision tree, and random forest show better results in terms of precision, recall and F1 for 1-day-ahead forecasting when these algorithms trained with the data of Shell Company. While, for testing data, we observe that extra tree shows low values in terms of precision, recall and F1. After and based on the prediction of these three companies we come to understand that the logistic regression outperforms the other machine learning models.

6. CONCLUSION

Forecasting the course of stock prices is significantly enough for having the development of successful share trading strategies. Guaranteeing attractive benefits for the investors can take place if there is effective prediction of closing stock prices. Machine learning algorithms have the ability to process and forecast almost reliable closing prices for historical stock patterns. In this study, we study the portfolio of three different companies belong to the energy sector, and the aim is to compute, using historical data, the opening price of next day stock. Eight different Machine Learning are applied on the data and are evaluated using several performances metrics to reach the aim and complete the task. The majority of research in this field is mostly interested in determining if the market price will rise or fall in the future.

The findings indicate that all regression and classifier Machine Learning algorithms are capable of forecasting the oil price with some differences. Furthermore, it is difficult to choose the best model among the classifier machine models for the analyzed companies; however, forecasts based on the logistic machine is often the most accurate. While, in the class of the regression machine learning models, the results indicate the forecasting superiority of the linear regression and ridge regression. So, forecasting the oil price will provide more details regarding the potential behaviour of the energy stock price if it is forecasted with a certain degree of precision. As a result, this strongly sends a signal to investors that demand forecasts may be done using computational intelligence techniques. Showing that stock values in the past can be used to forecast future prices is another goal this research aims at. According to the findings, past stock markets provide details that can be used to forecast future prices.

This attempt can be developed into several studies and direction in the future. First of all, other machine learning methods like neural networks or hybrid models can be applied. Second, other proxies of volatility, like the realized variance or the bi-power variation, can be used in sake of evaluating the forecasts. Then, the comparison of the models can be performed for simulated data assuming different generating processes.

REFERENCES

Abu-Mostafa, Y.S., Atiya, A.F. (1996), Introduction to financial forecasting. Applied Intelligence, 6, 205-213.

Bai, Y., Li, X., Yu, H., Jia, S. (2021), Crude oil price forecasting incorporating news text. International Journal of Forecasting, 38, 367-383.

Binrong, W., Lin, W., Sheng-Xiang, L., Yu-Rong, Z. (2021), Effective crude oil price forecasting using new text-based and big-data-driven model. Measurement, 168, 108468.

Bustos, O., Pomares-Quimbaya, A. (2020), Stock market movement forecast: A systematic Review. Expert Systems with Applications, 156, 113464.

Butler, S., Kokoszka, P., Miao, H., Shang, H.L. (2021), Neural network prediction of crude oil futures using B-splines. Energy Economics, 94, 105080.

Chen, S. (2002), Genetic Algorithms and Genetic Programming in Computational Finance. Amsterdam: Kluwer Academic Publishers.

Christodoulos, C., Michalakelis, C., Varoutas, D. (2010), Forecasting with limited data: Combining ARIMA and diffusion models. Technological Forecasting and Social Change, 77, 558-565.

Cohen, G. (2020), Forecasting bitcoin trends using algorithmic learning systems. Entropy, 22(8), 838.

Dehuri, S., Cho, S.B. (2008), Classification rule mining using gene expression programming. In: Third International Conference on Convergence and Hybrid Information Technology. Vol. Piscataway: IEEE. p2754-2760.

Dowling, M., Cummins, M., Lucey, B.M. (2014), Psychological barriers in oil futures markets. Energy Economics, 53, 293-304.

Fan, D., Sun, H., Yao, J., Zhang, K., Yan, X., Sun, Z. (2021), Well production forecasting based on arima-lstm model considering manual operations. Energy, 220, 119708.

Hamilton, J.D. (2003), What is an oil shock? Journal of Econometrics, 113(2), 363-398.

Hooker, M.A. (1996), What happened to the oil price-macroeconomy relationship? Journal of Monetary Economics, 38, 195-213.

Hooman, A. (2020), A novel hybrid model for forecasting crude oil price based on time series decomposition. Applied Energy, 267, 115035.

Hooman, A., Seyed, B.E. (2020), A new hybrid model for forecasting brent crude oil price. Energy, 200, 117520.

Jammazi, R., Aloui, C. (2012), Crude oil price forecasting: Experimental evidence from wavelet decomposition and neural network modeling. Energy Economics, 34, 828-841.

Kadkhodaie-Ilkhchi, A., Rezaee, M., Rahimpour-Bonab, H. (2009), A committee neural network for prediction of normalized oil content from well log data: An example from South Pars Gas Field, Persian Gulf. Journal of Petroleum Science and Engineering, 65, 23-32.

Li, J., Xu, Z., Xu, H., Tang, L., Yu, L. (2017), Forecasting oil price trends with sentiment of online news articles. Asia-Pacific Journal of Operational Research, 34(2), 1740019.

Li, X., Shang, W., Wang, S. (2019), Text-based crude oil price forecasting: A deep learning approach. International Journal of Forecasting, 35, 1548-1560.

Lisboa, P.J.G., Vellido, A., Edisbury, B. (2000), Business applications of neural networks. In: Business Applications of Neural Networks: The State of-the-art of Real-world Applications. Singapore: World Scientific. p7-22.

Lopez, H.S., Weinert, W.R. (2004a), EGIPSYS: An enhanced gene expression programming approach for symbolic regression problems. International Journal of Applied Mathematics and Computer Science, 14, 375-384.

Lopez, H.S., Weinert, W.R. (2004b), EGIPSYS: An enhanced gene expression programming approach for symbolic regression problems. International Journal of Applied Mathematics and Computer Science, 14, 375-384.

Margny, M.H., El-Semman, I.E. (2005), Extracting logical classification rules with expression programming: Micro array case study. IN: Conference Proceedings, AIML 05. Cairo, Egypt.

Narayan, P.K., Narayan, S. (2014), Psychological oil price barrier and firm returns. Journal of Behavioral Finance, 15(4), 318-333.

Narayan, P.K., Sharma, S.S. (2014), Firm return volatility and economic gains: The role of oil prices. Economic Modelling, 38, 142-151.

Patel, J., Shah, S., Thakkar, P., Kotecha, K. (2015), Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.

Paz-Marín, M., Campoy-Muñoz, P., Hervás-Martínez, C. (2012), Non-linear multiclassifier model based on artificial intelligence to predict research and development performance in European countries. Technological Forecasting and Social Change, 79, 1731-1745.

Phan, D.H.B., Sharma, S.S., Narayan, P.K. (2015b), Stock return forecasting: Some new evidence. International Review of Financial

Analysis, 40, 38-51.
Pradhan, R.P., Arvin, M.B., Ghoshray, A. (2015), The dynamics of economic growth, oil prices, stock market depth, and other macroeconomic variables: Evidence from the G-20 countries. International Review of Financial Analysis, 39, 84-95.
Sermpinis, G., Laws, J., Karathanasopoulos, A., Dunis, C.L. (2012), Forecasting and trading the EUR/USD exchange rate with gene expression and Psi sigma neural networks. Expert Systems with Applications, 39, 8865-8877.
Shambora, W., Rossiter, R. (2007), Are there exploitable inefficiencies in the futures market for oil? Energy Economics, 29, 18-27.
Shobana, G., Umamaheswari, K. (2021), Forecasting by machine learning techniques and econometrics: A review. In: 2021 6th International Conference on Inventive Computation Technologies (ICICT). Piscataway: IEEE. p1010-1016.
Tang, Q., Chen, Y., Yang, H., Liu, M., Xiao, H., Wu, Z., Chen, H., Naqvi, S.R. (2020), Prediction of bio-oil yield and hydrogen contents based on machine learning method: Effect of biomass compositions and pyrolysis conditions. Energy and Fuels, 34(9), 11050-11060.
Tseng, F.M., Yu, H.C., Tzeng, G.H. (2002), Combining neural network model with seasonal time series ARIMA model. Technological Forecasting and Social Change, 69, 71-87.
Ullah, Z., Khan, M., Raza Naqvi, S., Farooq, W., Yang, H., Wang, S., Vo, D.V.N. (2021), A comparative study of machine learning methods for bio-oil yield prediction - A genetic algorithm-based features selection. Bioresource Technology, 335, 125292.
Varian, H.R. (2014), Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2), 3-28.
Wang, J., Tao, Q. (2008), Machine learning: The state of the art. IEEE Intelligent Systems, 23(6), 49-55.
Wang, Q., Jiao, W., Wang, P., Zhang, Y. (2021), A tutorial on deep learning-based data analytics in manufacturing through a welding case study. Journal of Manufacturing Processes, 63, 2-13.
Watkins, G.C., Plourde, A. (1994), How volatile are crude oil prices? OPEC Review, 18, 220-245.
Weigend, A.S., Gershenfeld, N.A. (1994), Time Series Prediction: Forecasting the Future and Understanding the Past. Reading, MA: Addison-Wesley.
Wu, J., Chen, Y., Zhou, T., Li, T. (2019), An adaptive hybrid learning paradigm integrating ceemd, arima and sbl for crude oil price forecasting. Energies, 12(7), 1239.
Wu, Y.X., Wu, QB., Zhu, J.Q. (2019), Improved eemd-based crude oil price forecasting using lstm networks. Physica A: Statistical Mechanics and its Applications, 516, 114-124.
Yu, L., Wang, S., Lai, K.K. (2008), Forecasting crude oil price with an emd-based neural network ensemble learning paradigm. Energy Economics, 30(5), 2623-2635.
Yu, L., Zhang, X., Wang, S. (2017), Assessing potentiality of support vector machine method in crude oil price forecasting. Eurasia Journal of Mathematics, Science and Technology Education, 13(12), 7893-7904.
Zhang, X., Lai, K.K., Wang, S.Y. (2008), A new approach for crude oil price analysis based on empirical mode decomposition. Energy Economics, 30(3), 905-918.
Zhao, L.T., Zeng, G.R., Wang, W.J., Zhang, Z.G. (2019), Forecasting oil price using web-based sentiment analysis. Energies, 12(22), 4291.
Zhao, Y, Li, J, Yu, L. (2017), A deep learning ensemble approach for crude oil price forecasting. Energy Economics, 66, 9-16.