Commodity Price Forecasting in the International Market: Using a Proposed Ensemble Approach, Time Series and Machine Learning Models

Disha Rajesh Ghosh [1]

[1]*Student, Indian Institute of Science Education and Research (IISER Bhopal), Bhopal (MP), India.*

Corresponding Author: dishaghosh.applications@gmail.com

Abstract: - This research paper presents a comprehensive analysis of regression and time-series models for predicting commodity prices in the international market, focusing on Brent Oil, US Soybeans, and US Wheat. The study evaluates the accuracy and effectiveness of these models using performance metrics such as Root Mean Squared Error (RMSE) and R2 score. Additionally, a novel hybrid model is proposed, incorporating Random Forest, Decision Tree, Gradient Boosting, and further refined using a meta model - Linear Regression. The results of the analysis indicate that the hybrid model outperforms the majority of the traditional models and time-series models in terms of forecasting accuracy. Time-series models, specifically ARIMA and Prophet, demonstrate impressive performance in in-sample prediction. However, challenges were encountered with the LSTM model, which exhibited it to be computationally intensive and required careful parameter selection. To address these limitations, the research proposes potential avenues for improvement, although it acknowledges that the accuracy of predictions cannot be guaranteed. The findings of this study provide valuable insights for policymakers, investors, and risk analysts, offering a deeper understanding of the performance of different models in predicting commodity prices.

Key Words: — ***Commodity price forecasting, regression models, time-series models, hybrid approach, performance evaluation.***

I. INTRODUCTION

Commodity prices play a crucial role in the economy, influencing various aspects of business and financial decision-making in the international market. The fluctuations in commodity prices have a significant impact on economic growth, inflation rates, and trade balances at both the global and national levels. Industries involved in commodity production, such as agriculture, energy, and mining, closely monitor price trends to guide their production and investment strategies. Moreover, businesses that rely on commodities as raw materials, including manufacturers, carefully track commodity prices to manage costs and plan their supply chains.

Manuscript revised July 31, 2023; accepted August 01, 2023. Date of publication August 04, 2023.

This paper available online at www.ijprse.com

ISSN (Online): 2582-7898; SJIF: 5.59

Commodity prices also present investment opportunities, attracting investors who seek to capitalize on price movements through commodity trading or investments in commodity-related financial instruments. Additionally, fluctuations in commodity prices directly impact consumer spending patterns, influencing costs related to fuel, food, and other goods. Geopolitically, commodity prices hold significance, affecting political stability, international relations, and trade dynamics among countries.

Accurate predictions enable market participants to make informed decisions regarding investments, trading strategies, and risk management. In recent years, researchers have explored various approaches, including traditional time series models, machine learning algorithms, and ensemble techniques, to enhance the accuracy and reliability of commodity price forecasts.

The period from January 1, 2000, to December 31, 2022, witnessed significant global events that had a profound impact on commodity markets. These incidents disrupted market dynamics, created economic uncertainties, and influenced the

behavior of commodity prices. Four key events, namely the 9/11 terrorist attacks, the 2008 financial crisis, the Russia-Ukraine war, and the COVID-19 pandemic, stood out during this period, leaving lasting effects on commodity markets. [8]

Keeping the most recent disruptive event – the COVID-19 pandemic, in view, which emerged in late 2019 and spread globally in 2020. The pandemic caused unprecedented disruptions to global economies, international trade, and commodity markets. Lockdown measures, supply chain disruptions, and shifts in consumer demand patterns significantly affected commodity prices across different sectors. [29]

Considering all the aforementioned reasons, this research is narrowed down to three major commodities in the international market – Brent Oil, Soybeans and Wheat. Oil price prediction holds significance as it is a crucial energy source that affects industries, transportation, and consumer behavior worldwide. Fluctuations in oil prices directly influence production costs, inflation rates, and consumer spending patterns. Similarly, predicting wheat and soybean prices is vital as these commodities are essential in the agriculture and food sectors. Wheat is a staple food crop, and soybeans are widely used in animal feed, cooking oil, and various food products.

Therefore, considering the importance of accurately forecasting commodity prices amidst these significant events, this research paper proposes a hybrid approach that integrates traditional machine learning models, time series models including the newly developed Meta's Prophet, and machine learning algorithms.

In this study, we analyze and compare the performance of the proposed ensemble model, traditional time series models such as ARIMA [4], and various machine learning algorithms including Random Forest [10], Gradient Boosting [1], and Support Vector Regressor [2].

Through the results of the accuracy metrics and graphs' evaluation, we assess the effectiveness of these models in forecasting commodity prices in the aftermath of the aforementioned incidents.

By developing an enhanced forecasting approach, we strive to contribute to the overall efficiency and stability of commodity markets in the international arena.

The subsequent sections of the paper are organized as follows. In Section 2, we present a comprehensive literature review, highlighting previous research conducted in this specific domain. Afterward, we delve into the data and methodology employed for our study, along with the accuracy metrics utilized to assess the performance of the algorithms. Finally, we present the outcomes and results obtained from our analyses.

II. Literature Review

Among the major crops, wheat has a prominent role in consumption in Indian households. However, the wheat production does not suffice to meet the needs of the growing population in India. Comprising 82% of farmers and their economic contributions, agribusiness is the primary source of income for 70% of rural households in India [5]. Apart from wheat, another commodity which is one of the important commodities, Oil, plays an increasingly significant role in the world economy since nearly two-thirds of the World's energy consumption comes from crude oil and natural gas [19]. Global oil prices have been increasing drastically which can result in uncertainty in the global economy. The crude oil price is basically determined by its supply and demand but is more strongly influenced by many irregular past/present/future events like weather, stock levels, GDP growth, political aspects, and so on [19]. Similarly, Soyabean is another important commodity that has a huge impact on the world economy. The United States of America and Brazil are the largest producers of Soybeans [7].

Quantitative methods, such as regression and time-series analytics, tend to be more systematic and dependable [1]. For the forecasting of crude oil prices, [19] proposed a new method based on a support vector machine (SVM). The methods are compared to the traditional ARIMA technique and Back Propagation Neural Network (BPNN). The experimental results showed that SVM outperformed the other two time-series methods.

Another method proposed by [3] was the use of the econometric model - Granger Causality Test to establish the relationship between two time-series data. In addition, the use of the multivariate - LSTM method paired with the Gradient descent optimization method proved to be more accurate than other forecasting methods.

According to the LSTM model constructed by [4] and its comparison to the traditional ARIMA and ANN models, the results show that, first, the LSTM model has a strong generalization ability, with stable applicability in forecasting crude oil prices with different timescales. Second, as compared to other models, the LSTM model generally has higher forecasting accuracy for crude oil prices with different

timescales. Third, an LSTM model-derived shorter forecast price timescale corresponds to lower forecasting accuracy.

Apart from the common commodities such as crude oil, wheat and soybeans, [24] performed prediction on the most infamous cash crop – cotton. They have employed LSTM and ARIMA models on the dataset and proved that the latter algorithm has outperformed the LSTM model. Moreover, the paper suggests working on the averaging forecasting method which could yield a better result in either method.

Considering the idea of forecasting, which is important not only in manufacturing and production but also in retail, [1] compared the proposed hybrid model of Random Forest, XGBoost and Linear Regression to the traditional regression models of machine learning, which included, Random Forest, XGBoost, gradient boost, Adaptive Boost, Artificial Neural Network. The proposed hybrid was the best model when measured against other methods using the various performance metrics.

As per the forecasting wheat price report by [18], traditional ARIMA was used. The model was used on the major producing wheat states and at a national level for twelve months of the year. The autocorrelation function and partial autocorrelation function were used to decide the best fit ARIMA model.

[17] has worked on wheat yield prediction using XGBoost, Decision Tree and Random Forest. The features used were the climatic conditions of the place which include rainfall, season, area of production, and state. The accuracy was compared based on the performance metrics.

Soyabean and its derivative Soybean Oil price forecasting were conducted by [4]. Their method explored Nonlinear Autoregressive Neural Network (NARNN) and its derivative Exogenous Nonlinear Autoregressive Neural Network (XNARNN). The usefulness of the machine learning approach for price forecasting issues of the two commodities is demonstrated, as well as the potential usefulness of prices of closely related commodities for providing predictive content.

[9] used Meta's Prophet for Time series prediction of temperature in Myitkyina, Myanmar. Their research was evidence that the Time series model – Prophet can be a better alternative to the conventional meteorological methods.

An automated agricultural price prediction system [25] was designed which is focused on finding a suitable model with suitable parameters and overcoming challenges like nonlinearity and the curse of dimensionality. They have worked on famous algorithms which involve the advanced RNN algorithm – LSTM, SVR, Prophet, XGBoost and ARIMA. They have worked on the dataset of Malaysia.

The price of Gold is volatile and depends on various economic factors and is one the most important commodity because it is considered an investment and kept as a reserve by all the nation's banks around the globe. [27] has worked on the "Gold price prediction". They have used machine learning algorithms such as Random Forest regressor, Decision Tree, SVR, linear regression, and Artificial Neural Network.

III. Data And Methodology

3.1 Dataset Description

Data Source: https://www.investing.com/

The dataset consists of seven columns representing various attributes of the commodities:

Date: This column records the weekly time frame, covering the period from January 1st, 2000 to December

31st, 2022.

Price: Reflects the weekly price of the commodity.

Open: Indicates the opening weekly price of the commodity.

High: Represents the highest price recorded during the respective week.

Low: Represents the lowest price recorded during the respective week.

Volume: Denotes the average volume of the commodity for the given week.

Change%: Represents the percentage change in the price of the commodity.

3.2 Data Partitioning

For analysis, the dataset was divided into two parts: a training dataset spanning from January 1st, 2000 to December 31st, 2019, and a testing dataset encompassing the period from January 1st, 2020 to December 31st, 2022.

3.3 Significance and Objective

By examining the price fluctuations of Brent Oil, US Soybeans, and US Wheat, we aim to gain insights into the complex dynamics of the global market. Specifically incorporating global incidents like the 9/11 terror attacks, the 2008 Financial Crisis, the COVID-19 pandemic, and the Russia-Ukraine war into commodity price prediction is essential for several reasons.

First and foremost, these events have had profound and lasting effects on the international market, causing fluctuations in commodity prices. In addition, by considering these events, our price prediction models can better capture the complexities and uncertainties in the global commodity market, leading to more accurate and robust forecasts.

IV. METHODOLOGY

4.1 Architecture of the entire analysis

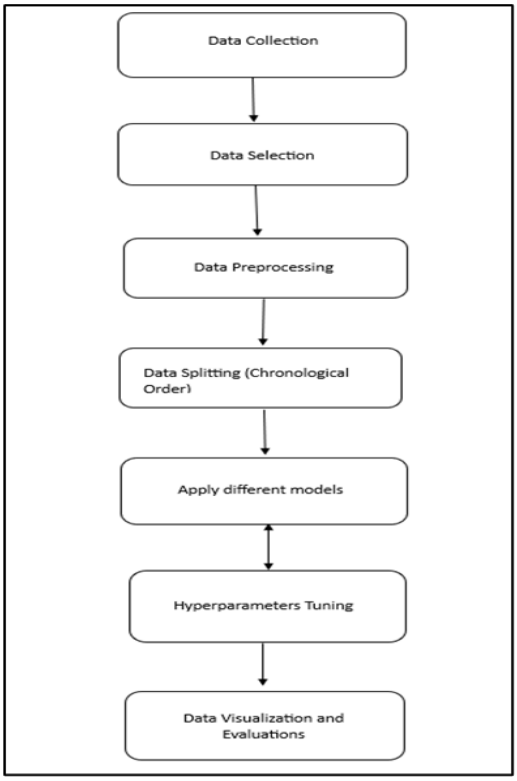

Fig.1. Flowchart Of the Methodology Used

During the data cleaning process, we replaced any null values with the mean of the respective column. Additionally, we carefully converted the data types of the attributes to their appropriate formats, ensuring that the data was structured optimally for accurate prediction.

To gain a deeper understanding of the dataset, we explored the relation of the "date" column and other attributes as shown in Figure.1. This analysis provided valuable insights into the relationships, trends, patterns, fluctuations and dependencies between different variables, enhancing our comprehension of the dataset's dynamics.

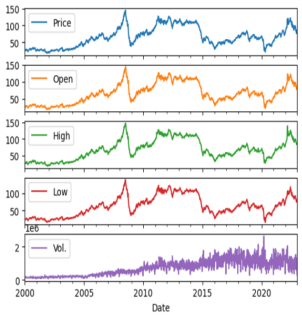

Fig. a. 1. Brent Oil attributes relation overtime

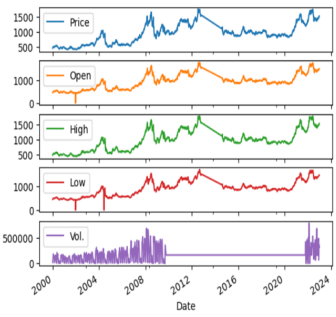

Fig. a. 2. US Soybeans attributes relation overtime

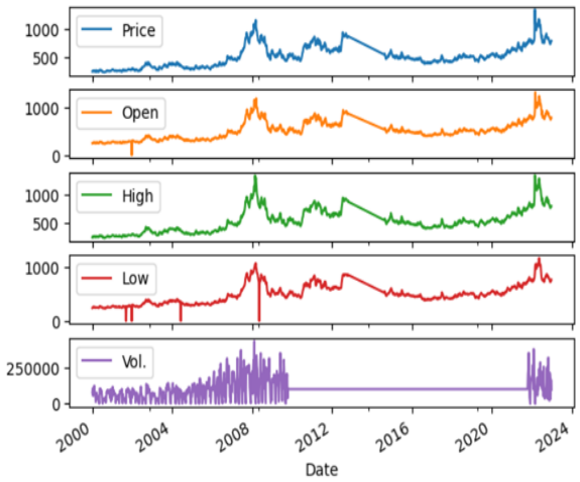

Fig. a. 3. US Wheat attributes relation overtime

Fig.2. a (1-3). Shows the relation of other attributes of Brent oil, US Soybeans and US Wheat with Date (2000 – 2023), respectively

The presence of a straight line in the Volume relation with Time for US Soybean and US Wheat from approximately 2010 to 2022 suggests the possibility of a lag or indicates that certain features may not have been adequately captured in the analysis. This observation raises the need for further investigation into potential time delays or missing factors that could affect the relationship between Volume and Date during this specific period.

Furthermore, we employed statistical measures of dispersion and quartile ranges to gain a comprehensive overview of the entire dataset as shown in the statistical summary table below.

By examining these statistical metrics, we obtained a better understanding of the spread and distribution of the data, enabling us to identify any potential outliers or anomalies.

Table.1. Statistical Summary for Brent Oil

	Price	Open	High	Low	Vol.
Mean	65.895942	65.839650	67.924900	63.574717	7.38130
Standard Dev.	29.469753	29.481392	30.081627	28.719811	4.89959
Minimum	17.750000	17.400000	18.700000	15.980000	6.00600
Maximum	144.490000	144.400000	147.500000	139.540000	2.6400
1st Quartile (Q1)	43.130000	42.955000	44.897500	40.435000	2.5682
2nd Quartile (Q2)	63.100000	62.940000	64.870000	60.960000	6.6941
3rd Quartile (Q3)	86.152500	86.350000	89.317500	83.545000	1.1125

Table.2. Statistical Summary for US Soybean

	Price	Open	High	Low	Vol.
Mean	925.6255	923.974111	948.017687	901.644885	168395.00
Standard Dev.	322.8709	323.264397	331.580002	314.104187	144296.86
Minimum	422.750	0.000000	426.000000	0.000000	840.00000
Maximum	1758.38	1772.630000	1788.880000	1725.630000	800250.000000
1st Quartile (Q1)	612.000	612.7500	629.000000	598.500000	52835.000
2nd Quartile (Q2)	919.12000	920.000000	941.380000	901.120000	138360.000000
3rd Quartile (Q3)	1100.6200	1094.000000	1140.000000	1053.000000	241575.000000

Table.3. Statistical Summary for US Wheat

	Price	Open	High	Low	Vol.
Mean	510.476559	510.985216	529.382631	492.447792	100985.661376
Standard Dev.	192.748542	194.404552	205.563651	184.894417	80118.306411
Minimum	236.750	0.000000	239.750000	0.000000	120.000000
Maximum	1348.00	1311.250000	1340.000000	1168.000000	431570.000000
1st Quartile (Q1)	356.50	355.250000	366.500000	344.500000	39475.0000
2nd Quartile (Q2)	486.13	486.380000	501.000000	471.250000	83190.000000
3rd Quartile (Q3)	611.750	611.440000	634.575000	592.190000	144255.000

4.2 Models and Parameters

The models selected are regression models, namely, Linear Regression, Random Forest Regressor, Gradient Boost Regressor, Decision Tree, Support Vector Machine, Polynomial Regression, and Neural Network models, ARIMA, LSTM and Meta's Prophet.

Apart from the traditional models, in our proposed approach, we have incorporated an Ensemble model which includes concepts like Bagging Ensemble, Stacking Ensemble, and Gradient Boost Ensemble. These combined models leverage their strength to increase their predictive power.

4.3 Linear Regression

Linear regression [30] is a statistical method for modelling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent and dependent variables and aims to find the best-fit line that can predict the dependent variable based on the independent variables. Mathematically, linear regression is represented as:

$$\boldsymbol{y = a_0 + a_1 x + \varepsilon}$$

Where, y = Dependent Variable (Target Variable)

x = Independent Variable (predictor Variable)

a_0 = intercept of the line (Gives an additional degree of freedom)

a_1 = Linear regression coefficient (scale factor to each input value).

ε = random error

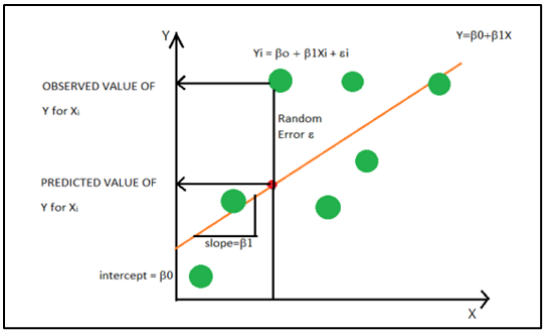

Fig.3. Graphical Representation of Linear Regression

4.4 Polynomial Regression

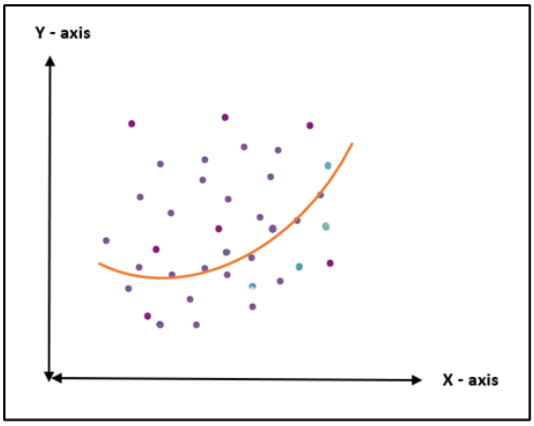

Fig.4. Graphical Representation of Polynomial Regression

Polynomial regression is utilized when dealing with non-linear relationships. It involves modelling the connection between the dependent variable and the independent variable using a polynomial function of degree 'n'.

The formula for polynomial regression can be expressed as:

$$\boldsymbol{y = a_0 + a_1 x_1 + a_2 x_0^2 \ldots \ldots a_n x_1^n}$$

In this equation, 'n' represents the input value, and the coefficients (a_0, a_1, a_2, ..., a_n) determine the shape and curve of the polynomial function.

4.5 Random Forest Regressor

Random Forest Regression [1] is a type of ensemble learning algorithm that uses multiple decision trees to predict the output variable. It is a powerful method for regression analysis as it can handle both categorical and continuous data, and can capture complex non-linear relationships between the independent and dependent variables.

Random Forest Regression works by building a large number of decision trees and combining their predictions to obtain a final prediction. Each decision tree is built on a random sample of the training data, and at each split in the tree, a random subset of features is considered. This randomness helps to prevent overfitting and ensures that the trees are diverse.

When making a prediction, each tree in the forest independently predicts the target variable based on the input features. The final prediction is then obtained by averaging the predictions of all the trees, in a process called Bagging. This ensemble approach helps to reduce the variance of the model and improve its predictive accuracy. This can be used in all forms of data, like categorical, continuous, or non-linear data.

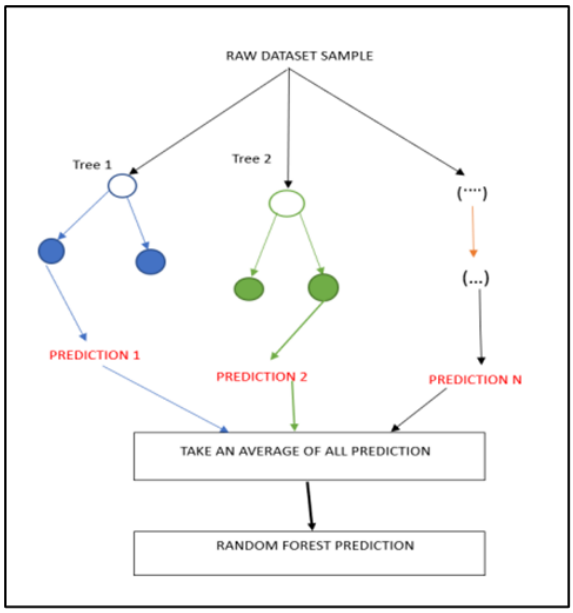

Fig.5. Graphical Representation of Random Forest [27]

The mathematics of Random Forest Regressor is explained below [10].

Scikit-learn determines the Gini Importance of each node for each decision tree, assuming that there are only two child nodes (binary tree):

Formula:

$$Ni_j = W_i C_j - W_{left(j)} C_{left(j)} - W_{right(j)} C_{right(j)}$$

Where, Ni sub(j)= the importance of node j

W sub(j) = weighted number of samples reaching node j

C sub(j)= the impurity value of node j

left(j) = child node from left split on node j

right(j)= child node from right split on node j

After calculating importance of each node, importance of the overall decision tree is calculated as follows:

$$Fi_i = \frac{(\sum \text{j: node j splits on feature i ni j})}{(\sum k\epsilon \text{ all features fik})}$$

Where, Fi sub(i)= the importance of feature i

ni sub(j)= the importance of node j

It is then normalized using the formula

$$Norm\ Fi_j = \frac{fi_j}{\sum_{j\epsilon\ all\ features} fi_j}$$

At the Random Forest level, the final feature relevance is determined by its average over all trees. Calculating the relevance of each attribute for each tree, then dividing that amount by the number of trees, yields:

$$RF\ Fi_j = \frac{\sum_{j\epsilon\ all\ trees} norm\ Fi_{ij}}{T}$$

Where, RF fi sub(i)= the importance of feature i calculated from all trees in the Random Forest model

Norm fi sub(ij)= the normalized feature importance for i in tree j

T = total number of trees

4.6 Decision tree

The decision tree [26] is a powerful tree-based structure used for accurate prediction in regression problems. By dividing the data into smaller and smaller subsets, decision trees offer enhanced predictive capabilities.

In decision tree regression, predictions are made by following a series of if-else conditions that split the data into subsets. Each split is based on a specific feature and threshold value. To make a prediction for a given data point, the path down the tree is

traversed until a leaf node is reached. At the leaf node, a specific value or the average value of the target variable is assigned as the prediction.

The prediction equation of the decision tree regressor is represented by a sequence of if-else conditions. It can be visualized as a hierarchical structure comprising branching nodes, with each node making a decision based on a feature and threshold value.

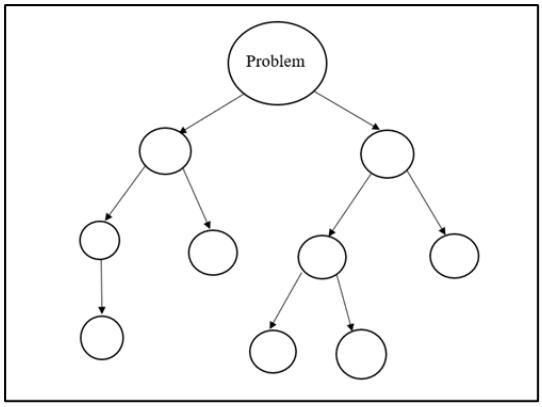

Fig.6. Pictorial Representation of Decision Tree [27]

4.7 Gradient Boost Regressor

The key idea behind Gradient Boosting Regression is to use the errors from the previous trees to train new trees that can better predict the target variable [1]. By doing this iteratively, the ensemble of decision trees can learn complex non-linear relationships between the input features and the target variable. There are 3 simple steps:

- Initialize a model (F0) to predict target variable y. An error (called residual) is produced.
- An alternate model h1 is prepared from this residual.
- F0 is added to h1 to give F1 (the boost). The error is reduced.
- The steps are repeated till the mean squared error between the models is minimum. Usually, the fixed number of iterations is selected based on the datasets.

The process of gradient boosting is done as follows [11]:

Let the training data be $(x_i, y_i)_{i=1}^{n}$, where x is the feature(s) and y is the prediction. Let the number of iterations be M.

An initial function $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L\,(yi, \gamma)$is initialized, where γ is a parameter specifying the minimum reduction in the loss function that must be achieved to continue splitting a tree node into child nodes and L is a loss function. If we take mean squared error as the loss function, the modified function is-

$$f_0 x = argmin_\gamma \sum_{i=1}^{n} (\gamma_i - \gamma)^2$$

Taking the first differential of $f_0\,(x)$, we get $\sum_{i=1}^{N} L\,(\frac{yi}{N})$which means this function minimizes the mean. A separate mean is calculated for the part of the dataset greater than the splitting means and less than or equal to the mean. This is our $h_1(x)$.

The boosted function $f_1(x)$ will simply be:

$$f_1(x) = f_0(x) + h_1(x)$$

These steps are repeated M times, or till the mean squared error is minimum.

4.8 Support Vector Regressor

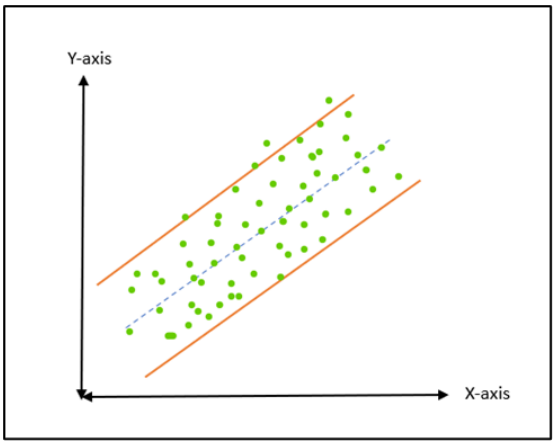

Fig.7. Pictorial Representation of Support Vector Regressor [23]

The red line represents the decision boundary, blue dotted line represents the hyperplane and the green dots are the data points. [2]

Let's consider two lines drawn at a distance of 'a' from the hyperplane, one above and one below. These lines represent the decision boundaries. The variable 'a' is epsilon, which determines the width of the margin.

Assuming the equation of the hyperplane as Y = wx + b, the decision boundary equations can be expressed as:

wx + b = +a (i)

wx + b = -a (ii)

To satisfy the SVR, any hyperplane should meet the condition:

-a < Y - wx + b < +a (iii)

In other words, we only consider data points that fall within the decision boundary and have a minimal error or lie within the Margin of Tolerance. This approach allows us to create a more accurate model that fits the data well.

4.9 ARIMA

ARIMA [2][4], or Autoregressive Integrated Moving Average, is a widely-used time-series forecasting method that models the relationship between a dependent variable and its past values, as well as the errors of past predictions, to make future predictions. The method is powerful, and flexible, and can be used to model and forecast a wide range of time-series data. [4]

The ARIMA model has three key components: autoregression (AR), integration (I), and moving average (MA). The AR component models the relationship between the dependent variable and its past values, while the MA component models the relationship between the errors of past predictions and the current prediction. The I component specifies the order of differencing needed to make the time series stationary.

The mathematical model for ARIMA (p, d, q) is given by: [14]

$$\boldsymbol{y(t) = c + \varphi_1 * y_{(t-1)} + \varphi_2 * y_{(t-2)} \cdots . \varphi_p * y_{(t-p)} + \theta_1 * \varepsilon_{(t-1)} + \theta_2 * \varepsilon_{(t-2)} + \cdots \theta_q * \varepsilon_{(t-q)} + \varepsilon(t)}$$

Where, y(t) is the value of the dependent variable at time t

c is a constant or intercept term

$\varphi_1, \varphi_2 \ldots . \varphi_p$ are the autoregressive coefficients

ε(t) is the error term at time t

$\theta_1, \theta_2 \ldots . \theta_q$ are the moving average coefficients

p, d, and q are the orders of the AR, I, and MA components, respectively.

The AR component is represented by the autoregressive terms, while the MA component is represented by the moving average terms. The 'I' component is represented by the differencing operator (1-B)^d, where B is the backshift operator.

The ARIMA model is fitted to the data using maximum likelihood estimation, which involves finding the parameter values that maximize the likelihood of the observed data given the model. Once the model is fitted, it can be used to make predictions for future time points.

In order to ensure the adequacy of ARIMA(p,d,q) model, the time series data was examined for stationarity using the Augmented Dickey-Fuller unit root test. [14]

However, there are some limitations to the ARIMA model.

- It assumes that the data is stationary, and it can be difficult to determine the appropriate order of differencing, AR, and MA terms. [14].
- It does not capture complex patterns or trends in the data that are not captured by past values or past errors. [14]
  p, d, and q are the orders of the AR, I, and MA components, respectively.

The AR component is represented by the autoregressive terms, while the MA component is represented by the moving average terms. The 'I' component is represented by the differencing operator $(1 - B)^d$.

Where, B is the backshift operator.

The ARIMA model is fitted to the data using maximum likelihood estimation, which involves finding the parameter values that maximize the likelihood of the observed data given the model. Once the model is fitted, it can be used to make predictions for future time points.

4.10 Long – Short term Memory (LSTM)

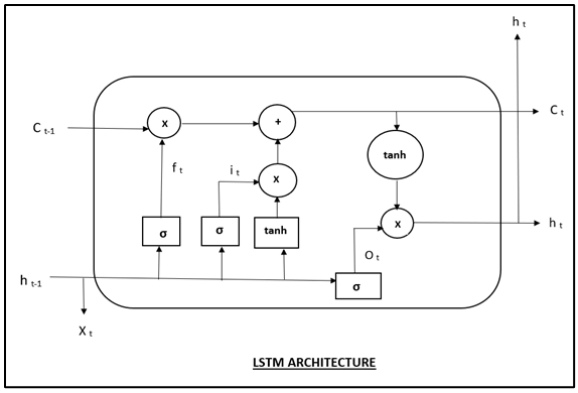

Fig.8. Architecture of LSTM [15]

LSTM [3][4][8], a specialized type of recurrent neural network (RNN), tackles problems by addressing the issue of vanishing gradients. [14]

In an LSTM module, there exists a cell state and three gates that enable selective learning, retention, or removal of information from each unit. The cell state facilitates the smooth flow of information through the units, ensuring minimal alterations by permitting limited linear interactions. Each unit encompasses an input gate, output gate, and forget gate, which controls the

addition or removal of information to the cell state. The forget gate, employing a sigmoid function, determines which information from the previous cell state should be disregarded. The input gate, through a pointwise multiplication of 'sigmoid' and 'tanh' operations, regulates the information flow to the current cell state. Finally, the output gate determines the information to be passed on to the subsequent hidden state.

Mathematically, the procedure for LSTM is expressed as follows: [8]

$$f_t = \alpha(w_t[h_{t-1}, x_t] + b_y) \quad (1)$$

$$i_t = \alpha(w_i[h_{t-1}, x_t] + b_i) \quad (2)$$

$$o_t = \alpha(w_o[h_{t-1}, x_t] + b_o) \quad (3)$$

$$C_t = f_f C_{t-1} + i_t * t_h(w_c[h_{t-1}, x_t] + b_c) \quad (4)$$

$$h_t = O_t * \beta(C_t) \quad (5)$$

Moreover, LSTM models possess exceptional memory capabilities, both in the long-term and short-term, which prevent significant loss of historical state information regarding crude oil prices. These models effectively extract historical data on commodity prices, considering current data characteristics. LSTM models excel in uncovering long-term dependencies in commodity price sequence data. They also possess the ability to automatically identify nonlinear features and complex patterns in commodity prices, yielding remarkable forecasting performance. As LSTM stands as a powerful prediction tool, it has found widespread use in prediction-related domains. Hence, to enhance the accuracy of crude oil price forecasting, we have selected the LSTM model for this study. [4]

4.11 Meta's Prophet Model

Meta's Prophet model [20] is a univariate time-series model developed by Meta, previously Facebook. It is typically modeled as normally distributed noise. It works using only two columns: ds (date time column) and y (target variable).

The equation given by the prophet is as follows:[9]

y(t) = g(t) + s(t) + h(t) + ε(t)

where ε(t) is error term

g(t)= trends

s(t)= seasonality

h(t)= holidays

4.12 Ensemble model (Random Forest – Decision Tree – Gradient Boost)

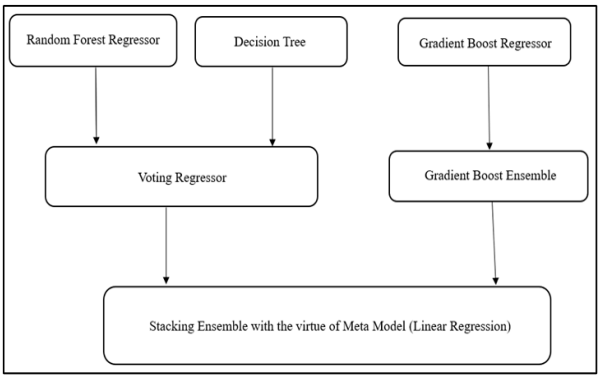

Fig.9. Architecture of Proposed Hybrid model

Hybrid models [1] are created by integrating multiple models to enhance accuracy, flexibility, and overall capability. By combining the strengths of different models, hybrid approaches aim to achieve superior performance in various applications and domains.

The development of a hybrid model is driven by two primary reasons: [1]

- Mitigating the risk of a single forecast making inaccurate predictions in certain specific conditions.
- Enhancing the overall performance by leveraging the strengths of individual models and overcoming their limitations.

In our research, we employed the Random Forest regressor and decision tree as base models, utilizing the bagging technique [17]. This approach involved creating multiple samples from the training data through bootstrapping. Each model learned from these samples and generated individual predictions.

To obtain the final prediction for regression, we employed the averaging method, where the predictions from all the models were combined.

In our ensemble model, we utilized the Gradient Boosting technique [1], which was complemented by a bagging approach incorporating the average results of Random Forest and Decision Tree models. Through a series of iterative iterations, our ensemble model effectively combined these diverse models, with a focus on continuously reducing the residual errors. This technique enabled the ensemble model to learn from the mistakes made by the previous models, leading to continuous improvement in performance.

To achieve the final prediction, the base models generated by the bagging technique and the iteratively improved models from Gradient Boosting were combined with a meta-model. In our research, we selected linear regression as the meta-model. The choice of linear regression was driven by the need for interpretability in the ensemble model. By employing linear regression as the meta-model, we were able to gain insights into the relative importance of each base model's prediction and understand their contributions to the overall prediction of the ensemble.

4.13 Accuracy Metrics

4.13.1 R2 Score

The R2 score ranges from 0 to 1, with 1 indicating a perfect fit of the model to the data, and 0 indicating that the model does not explain any of the variability in the data. A negative R2 score indicates that the model is worse than a horizontal line, which simply predicts the mean of the dependent variable for all observations.

The R2 score is calculated as the ratio of the explained variation to the total variation in the dependent variable. The formula for R2 score [1] is:

$$\boldsymbol{R^2 = 1 - (SS_{res}/SS_{tot})}$$

where, SS_{res} = sum of squared residuals (the difference between actual and predicted values)

SS_{tot} = total sum of squares (the difference between actual and mean values)

4.13.2 Root Mean Square Error (RMSE)

The formula to calculate RMSE is as follows: [2]

$$\sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}$$

where yt = represents the actual value

ŷt= predicted values

and, N= the number of testing data sets.

V. RESULTS AND ANALYSIS

The evaluation of regression and machine learning algorithms, such as Random Forest, Decision Tree, Gradient Boost, Support Vector Regressor, Linear Regression, and Polynomial Regression, included the calculation of the R2 score. However, for the time series models, namely LSTM, ARIMA, and Prophet, the R2 score was not calculated. This is because the R2 score is considered an in-sample metric [13], and it may not effectively capture the underlying trends and seasonality in time series data. Instead, the selection of the best fit model among the time series models was based on the lowest values of Mean Absolute Error (MAE) or Mean Squared Error (MSE) [12]. By prioritizing models with lower MAE or MSE, we can better account for the characteristics and performance of time series forecasting. [12] [13]

Table.4. Brent Oil Price Prediction

Model Name	RMSE Score	R2 Score
Random Forest Regressor	3.017	0.9855
Gradient Boost Regressor	2.1983	0.9923
Decision Tree	5.7174	0.9479
Linear Regression	2.0698	0.9931
Polynomial Regression	2.2177	0.9921
Support Vector Regressor	2.1192	0.9928
LSTM	3.5544	-
ARIMA	48.0175	-
Prophet	35.7286	-
Ensemble Model	2.1523	0.9926

Table.5. US Soybeans Price Prediction

Model Name	RMSE Score	R2 Score
Random Forest Regressor	44.6727	0.9716
Gradient Boost Regressor	34.8089	0.9827
Decision Tree	35.929	0.9816
Linear Regression	23.992	0.9918
Polynomial Regression	21.6867	0.9933
Support Vector Regressor	21.0677	0.9936
LSTM	59.8233	-
ARIMA	501.3926	-
Prophet	504.0936	-
Ensemble Model	31.8888	0.9855

Table.6. US Wheat Price Prediction

Model Name	RMSE Score	R2 Score
Random Forest Regressor	44.2007	0.9335
Gradient Boost Regressor	40.4812	0.9442
Decision Tree	47.0737	0.9246
Linear Regression	24.9834	0.9787
Polynomial Regression	44.2026	0.9335
Support Vector Regressor	20.1539	0.9861
LSTM	41.8746	-
ARIMA	327.7541	-
Prophet	279.915	-
Ensemble Model	38.7695	0.9488

Based on the analysis, Table.4.; for Brent Oil Price Prediction, the proposed ensemble model demonstrated the highest

accuracy, closely followed by the traditional Support Vector Regressor. The R2 scores achieved were 0.9926 and 0.9928, with corresponding RMSE scores of 2.153 and 2.1192, respectively.

Regarding the Table.5.; for US Soybeans Price Prediction, the proposed model outperformed most of the traditional regression and time series models. However, the top-performing models were Polynomial Regression and Support Vector Regressor, with R2 scores of 0.9933 and 0.9936, respectively.

The ensemble model achieved an R2 score of 0.9855, showing promising results compared to the best performing model in this dataset.

In the case of Table.6.; for US Wheat Price Prediction, the Support Vector Regressor demonstrated the best performance, achieving an R2 score of 0.9861 and an RMSE value of 20.1539. This indicates the superior predictive capability of the Support Vector Regressor in this particular dataset.

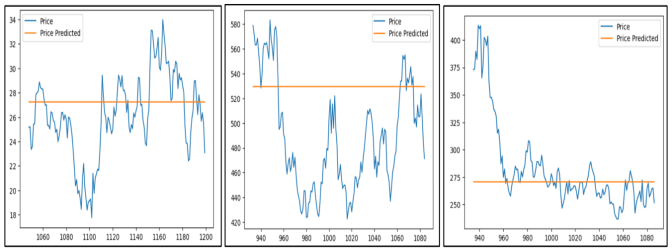

Fig.10. LSTM forecasting graphs for Brent oil, US Soybeans and US Wheat, respectively

There are several probable reasons for the unsatisfactory performance of the LSTM model:

- The LSTM model faces challenges in processing continuous input streams that are not segmented into explicit subsequences. Without defined endpoints to reset the network's internal state, the state can grow indefinitely, potentially leading to network breakdown [16].
- LSTM models are inherently more complex than traditional Recurrent Neural Networks (RNNs), requiring a larger amount of training data to accurately compute results. [16]
- Considering the dataset size, which contained over 1000 rows, training the LSTM model becomes time-consuming. However, it is important to note that the model needs sufficient time to learn the parameters within the LSTM cells, which can be computationally intensive. [16]

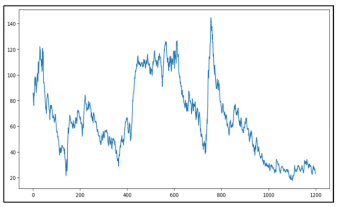

Fig. V. e. Brent Oil Price over the time indexes

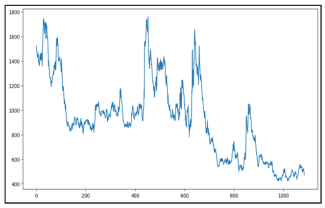

Fig. V. f. US Soybeans over the time indexes

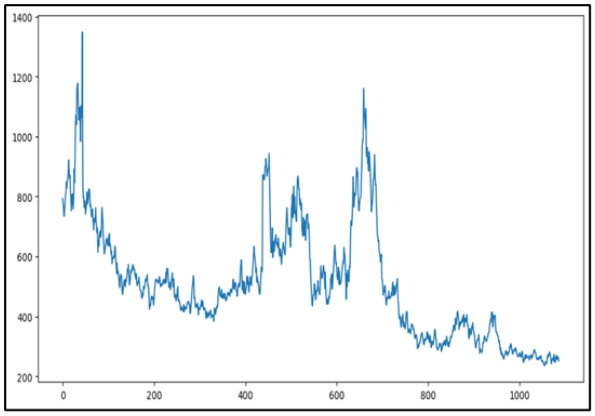

Fig.11. V. e-g. US Wheat over the time indexes

The graphs shown in Fig.11. V. e, Fig.11. V. f, Fig.11. V. g, represents the prices w.r.t to the indexes. The X - axis represents the indexes and the Y - axis represents the Prices.

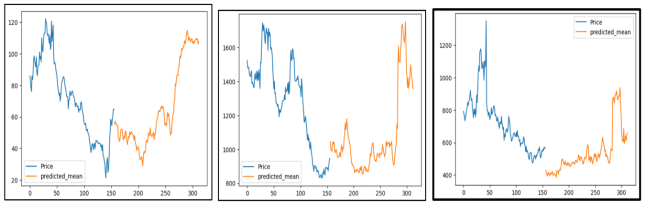

Fig.12. V. h. ARIMA test set forecasting graphs for Brent Oil, US Soybeans and US Wheat, respectively

In the Fig.12. V. h., X-axis represents the indexes and the Y-axis represents the prices.

When the indexes 0 to 300, represented in Fig.12. V. e, Fig.12. V. f, Fig.12. V. g, is compared to the indexes 0 to 300 in test graphs shown in Fig.12. V. h., we can see that the values predicted are almost accurate and satisfactory.

Among the models evaluated, ARIMA performed moderately in terms of predictive accuracy. Although its RMSE values were relatively higher compared to other models mentioned, it is important to consider the mean of the target variable. Greater the difference between the RMSE values and the mean suggests that the predictions made by ARIMA are still considered satisfactory, as they are within an acceptable range.

One thing in common which can be noticed in for all the datasets is that the ARIMA model has the highest RMSE value, prospective reasons could be: [14]

- It assumes that the data is stationary, and it can be difficult to determine the appropriate order of differencing, AR, and MA terms.
- It does not capture complex patterns or trends in the data that are not captured by past values or past errors.

In the Prophet's model, the x - axis represents the target variable (Price) and the y - axis represents the year. Univariate time series prediction is made to forecast the price of the previously mentioned commodities by learning the historic data which is used as an input to the prediction model.[9]

The prediction results revealed that the model is satisfactorily fitted to the historical data. Moreover, through developing the predictive model, it can be rightly said; Prophet Model is able to yield good results for price prediction of different commodities in the international market and can be used as an alternative to conventional econometric methods.[9]

VI. Conclusion

Through a comprehensive evaluation of the algorithms in a real-world context, the proposed ensemble model exhibited the ability to generate accurate weekly price predictions for various commodities in the international market.

In this study, a hybrid machine learning model combining XGBoost, Random Forest (RF), and Decision Tree (DT) has been proposed to predict commodity prices. The model leverages attributes such as opening price, lowest and highest price, and volume to enhance accuracy. Initially, the dataset was divided and trained on multiple gradient boosting models using the gradient boost ensemble technique. Additionally, decision tree and random forest models were trained using the bagging technique. The predictions from these models were combined to create a new dataset, which served as input for the meta model, Linear Regression (LR), to generate the final predictions.

The evaluation of the proposed hybrid model RF-DT-XGBoost-LR using RMSE and R2 score metrics indicates its superior performance compared to most machine learning models for Brent Oil (RMSE=2.153, R2=0.9926), US Soybean (RMSE=31.8888, R2=0.9855), and US Wheat (RMSE=38.7695, R2=0.9488). The results are also comparable to the best-performing traditional machine learning models like support vector regressor and linear regression. Additionally, satisfactory forecasts were achieved by the time series models ARIMA and Prophet, though slightly lower than the ensemble model. The high R-squared scores indicate that the model explains a significant portion of the data and variables, with percentages of 99.26%, 98.55%, and 94.88% for Brent Oil, US Soybeans, and US Wheat, respectively.

It is also important to note that the forecasting accuracy should be interpreted in the context of significant events that took place during the 22-year period, such as the 9/11 attacks, 2008 financial crisis, and the COVID-19 pandemic. These events might have influenced the accuracy of the forecasts and should be taken into consideration when analyzing the results.

While there may exist more advanced methods for brent oil, US Soybean and US Wheat price prediction, this study acknowledges the limitations imposed by time and personal resources, preventing the implementation of alternative approaches. For instance, the LSTM model, although beneficial with its long-term memory capability, it faces challenges such as a lengthy gradient path and time-consuming training process when dealing with input sequences of considerable length. Furthermore, the recursive nature of LSTM restricts the possibility of parallel training. Despite these constraints, the chosen models and methodologies have provided valuable insights and results in the context of the study's scope and available resources.

References

[1]. Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach. Operations Research Forum, 3(4).

[2]. Xie, W., Yu, L., Xu, S., & Wang, S. (2006). A new method for crude oil price forecasting based on support Vector Machines. Computational Science – ICCS 2006, 444–451.

[3]. Sen, A., Dutta Choudhury, K., & Kumar Datta, T. (2023). An analysis of crude oil prices in the last decade (2011-2020): With Deep Learning Approach. PLOS ONE, 18(3).

[4]. Zhang, K., & Hong, M. (2022). Forecasting crude oil price using LSTM Neural Networks. Data Science in Finance and Economics, 2(3), 163–180.

[5]. Nayana, B. M., Kumar, K. R., & Chesneau, C. (2022). Wheat yield prediction in India using principal component analysis-multivariate adaptive regression splines (PCA-Mars). AgriEngineering, 4(2), 461–474.

[6]. Xu, X., & Zhang, Y. (2022). Soybean and soybean oil price forecasting through the Nonlinear Autoregressive Neural Network (NARNN) and NARNN with exogenous inputs

(NARNN–X). Intelligent Systems with Applications, 13, 200061.

[7]. Colussi, J. and G. Schnitkey. "Brazil Likely to Remain World Leader in Soybean Production." farmdoc daily (11): 105, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, July 12, 2021.

[8]. Ben Ameur, H., Boubaker, S., Ftiti, Z., Louhichi, W., & Tissaoui, K. (2023). Forecasting commodity prices: Empirical evidence using Deep Learning Tools. Annals of Operations Research.

[9]. OO, Z. Z., & PHYU, S. (2020). Time series prediction based on Facebook Prophet: A case study, temperature forecasting in Myintkyina. International Journal of Applied Mathematics Electronics and Computers, 8(4), 263–267.

[10]. Hassouna, Fady. (2020). Re: What is the problem with using R-squared in time series models?.

[11]. Davydenko, Andrey. (2020). Re: What is the problem with using R-squared in time series models?.

[12]. Andreea-Cristina PETRICĂ & Stelian STANCU & Alexandru TINDECHE, 2016. "Limitation of ARIMA models in financial and monetary economics," Theoretical and Applied Economics, Asociatia Generala a Economistilor din Romania - AGER, Volume 0(4(609), W), pages 19-42, Winter.

[13]. Van Houdt, Greg & Mosquera, Carlos & Nápoles, Gonzalo. (2020). A Review on the Long Short-Term Memory Model. Artificial Intelligence Review. 53.

[14]. D'informatique, D & N, Ese & Esent, Pr & Au, Ee & Gers, Felix & Hersch, Prof & Esident, Pr & Frasconi, Prof. (2001). Long Short-Term Memory in Recurrent Neural Networks.

[15]. B M, Dr. S., N K, Dr. C., T, Dr. P., & R, Dr. R. (2022). Rice and wheat yield prediction in India using decision tree and Random Forest. Computational Intelligence and Machine Learning, 3(2), 1–8.

[16]. Darekar, Ashwini & Reddy, An Amarender. (2018). Forecasting Wheat Prices in India. 10. 54-60.

[17]. Xie, Wen & Yu, Lean & Xu, Shanying & Wang, Shouyang. (2006). A New Method for Crude Oil Price Forecasting Based on Support Vector Machines. 3994. 444-451.

[18]. Žunić, E., Korjenić, K., Hodžić, K., & Đonko, D. (2020). Application of facebook's prophet algorithm for successful sales forecasting based on real-world data. International Journal of Computer Science and Information Technology, 12(2), 23–36.

[19]. Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization, 1–25.

[20]. Grigoryan, H. (2017). Stock market trend prediction using support vector machines and variable selection methods. Proceedings of the 2017 International Conference on Applied Mathematics, Modelling and Statistics Application (AMMSA 2017).

[21]. Huang, S., Tian, L., Zhang, J., Chai, X., Wang, H., & Zhang, H. (2021). Support vector regression based on the particle swarm optimization algorithm for tight oil recovery prediction. ACS Omega, 6(47), 32142–32150.

[22]. Ly, R., Traore, F., & Dia, K. (2021). Forecasting Commodity Prices Using Long-Short-Term Memory Neural Networks.

[23]. Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K., & Liew, X. Y. (2021). Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques. Advances in Science, Technology and Engineering Systems Journal, 6(4), 376–384.

[24]. Baser, P. ., Jatinderkumar R. Saini, & Baser, N. . (2023). Gold Commodity Price Prediction Using Tree-based Prediction Models. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 90–96.

[25]. Tripurana, N., Kar, B., Chakravarty, S., Paikaray, B. K., & Satpathy, S. (2022). Gold Price Prediction Using Machine Learning Techniques. Workshop on Advances in Computation Intelligence, Its Concepts & Applications at ISIC 2022, May 17-19, Savannah, United State.

[26]. Tripurana, N., Kar, B., Chakravarty, S., Paikaray, B. K., & Satpathy, S. (2022). Gold Price Prediction Using Machine Learning Techniques. Workshop on Advances in Computation Intelligence, Its Concepts & Applications at ISIC 2022, May 17-19, Savannah, United State.

[27]. Raj, A., Mukherjee, A. A., de Sousa Jabbour, A. B. L., & Srivastava, S. K. (2022). Supply chain management during and post-COVID-19 pandemic: Mitigation strategies and practical lessons learned. Journal of business research, 142, 1125–1139.

[28]. Edelman ER, van Kuijk SMJ, Hamaekers AEW, de Korte MJM, van Merode GG and Buhre WFFA (2017) Improving the Prediction of Total Surgical Procedure Time Using Linear Regression Modeling. Front. Med. 4:85.