© International Journal of Scientific Research in Science and Technology
Print ISSN : 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)

Forecasting Commodity Prices using Machine Learning

Mr. Harsh Gupta*1, Mrs. Rashmi Kumari2, Mr. Swapnil Rajput3, Mrs. Nupur Puri4, Mrs. R Aafrein5

*[1]Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
harshgup11@gmail.com[1]
[2]Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
rashmioff0708@gmail.com[2]
[3]Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India
swapnilrajput977@gmail.com[3]
[4]Computer science and Engineering, Presidency University, Bangalore Karnataka, India
nupurpuri0214@gmail.com[4]
[5]Computer science and Engineering, Presidency University, Bangalore Karnataka, India
aafreinrafiq11@gmail.com[5]

ABSTRACT

This research endeavor focuses on forecasting price movements across 14 diverse commodities by leveraging technical analysis and comprehensive dataset correlations. The dataset integrates pivotal economic indicators and historical commodity prices, covering a wide spectrum from natural gas to agricultural products. Key technical indicators, such as lagged values, moving averages, MACD, historical volatility, and standard deviation, are used to predict percentage changes in commodity prices over time. The methodology underscores meticulous preprocessing steps, crafting technical features to capture intricate patterns and interrelationships within the dataset. These features are derived from a blend of time series analysis, statistical exploration, and machine learning techniques. An extensive suite of machine learning models is employed, including SVM, Decision Trees, KNN, XGBoost, Random Forests, Linear Regression, Gradient Boosting, VAR, VARIMA, VARMA, GRU, LSTM, and Extra Trees. Thorough evaluation with diverse metrics provides a comprehensive understanding of their efficacy in price prediction. Further enhancements include Isolation Forest for outlier detection, feature standardization with Standards Caler, and hierarchical clustering techniques, optimizing dataset analysis and model performance. Ultimately, this research aims to develop a precise predictive model merging technical analysis with economic indicators for commodity price forecasting. The potential insights could significantly impact decision-making in commodity trading and financial markets, presenting valuable contributions to this evolving landscape.

Keywords: Commodity forecasting, technical analysis, Economic indicators, Historical prices, Dataset correlations, Machine learning models, Price prediction, Time series analysis, Statistical exploration, Feature engineering, Outlier detection, Model evaluation, Decision-making in trading, financial markets, Predictive modeling.

I. INTRODUCTION

This research initiative represents a deep exploration into the convergence of data science with the ever-evolving dynamics of financial markets. Its primary objective is to architect an advanced predictive model capable of astutely analyzing and forecasting price fluctuations across a diverse array of 14 commodities. Anchored at its core is the utilization of historical data archives, intricate modeling techniques, and the nuanced correlations intrinsic to economic indicators, all meticulously orchestrated to navigate the complexities entrenched within commodity markets.

The expedition commences with a thorough scrutiny of historical datasets, each unveiling a distinct chapter in the narrative of market evolution. These datasets encapsulate the historical trajectories of various commodities across different timeframes, forming the foundational groundwork upon which advanced modeling techniques are meticulously applied. These techniques are tailored not merely for prediction, but to craft an intricately detailed and finely calibrated predictive landscape.

Acquiring a profound understanding of this landscape necessitates a holistic comprehension of the intricate interplay between economic indicators and the intricate dance of commodity prices. This intricate endeavor involves deciphering the multifaceted interactions among supply and demand dynamics, geopolitical shifts, and macroeconomic trends—a rich tapestry that shapes the undulating movements in commodity prices.

However, the ultimate aspiration transcends mere prediction; it strives to forge a dependable predictive model finely attuned to the dynamic nature of market forces. This journey signifies a fusion of insights gleaned from historical data, innovative methodologies in advanced modeling, and insightful analysis of correlations among economic indicators.

At its pinnacle lies the envisioned creation of an intuitive, real-time platform—a reservoir of invaluable insights guiding stakeholders through the complex pathways of commodity markets. It transcends the boundaries of a typical data repository; it stands as a guiding light, empowering decision-makers with actionable insights. This visionary platform transforms raw data into strategic foresight, empowering users to navigate the ebb and flow of commodity markets with confidence and precision.

II. METHODS AND MATERIAL

1. Data Collection:

The data for this study was sourced from multiple financial databases, including Bloomberg Terminal, Quandl, and Yahoo Finance. Historical records for 14 commodities, including Natural Gas, Gold, WTI Crude, Brent Crude, Soybeans, Corn, Copper, Aluminum, Zinc, Nickel, Wheat, Sugar, Coffee, and Cotton, were obtained. The dataset spanned 10 years, from January 2012 to December 2021. The criteria for selecting these commodities were based on their market significance and trading volumes.

2. Data Preprocessing:

Raw data underwent extensive preprocessing steps to ensure data quality and uniformity. Missing values were handled using forward and backward filling methods, and outliers were identified and corrected using robust statistical techniques. Additionally, the data was standardized to bring all features onto the same scale, preventing any bias in the modeling process.

3. Feature Engineering:

Several technical indicators were derived to enhance the dataset's predictive power. Lag features with intervals of 1, 3, 5, 9, 21, 100, and 200 days (about 6 and a half months) were calculated for each commodity price. Moving averages, including Simple Moving Averages (SMA) and Exponential Moving Averages (EMA) over periods of 5, 10, 20, 100, and 200 days (about 6 and a half months), were computed. Other features such as Moving Average Convergence Divergence (MACD), Historical Volatility (HV), and Standard Deviation (STD) were also incorporated into the dataset.

4. Model Selection and Training:

A variety of machine learning models were selected and trained on the preprocessed dataset. Models including Support Vector Machines (SVM), Decision Tree Regressors, K Neighbors Regressors, Linear Regression, Vector Autoregression (VAR), Vector Autoregression Moving-Average (VARMA), Vector Autoregression Integrated Moving-Average (VARIMA), and Long Short-Term Memory (LSTM) networks were implemented using Python's Scikit-learn and TensorFlow libraries.

5. Model Evaluation:

The performance of each model was assessed using key regression metrics. Metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared values were used to evaluate the models' accuracy, precision, and goodness of fit.

III. PROPOSED MEHODOLOGY

1. **Support Vector Machines (SVM):** Support Vector Machines are versatile models used for classification and regression tasks. SVMs aim to find an optimal hyperplane that best separates different classes or predicts continuous values. They work exceptionally well in high-dimensional spaces and are effective in cases where the data isn't linearly separable, thanks to kernel functions. SVMs are robust against overfitting and perform well with small to medium-sized datasets. However, they can be sensitive to the choice of kernel and parameters and might require significant computational resources for larger datasets.

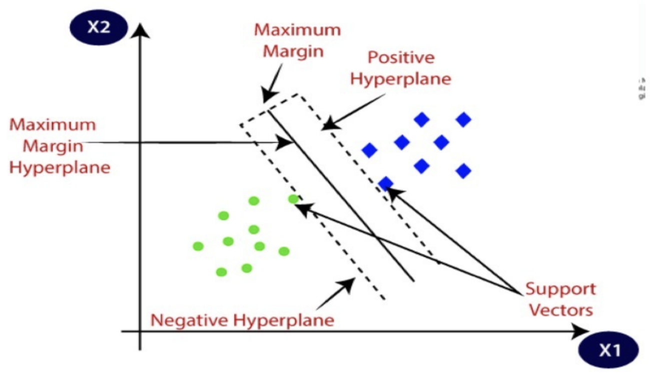


Figure 1- Support Vector Machines (SVM)

2. **Decision Trees:** Decision Trees are intuitive models that make decisions based on feature values. They recursively split the dataset into branches to create homogeneous subsets. These models are easy to interpret, handle both numerical and categorical data, and offer insights into feature importance. However, they're prone to overfitting, especially with deep trees.

Techniques like pruning or employing ensemble methods like Random Forests mitigate this issue and improve their generalization ability.

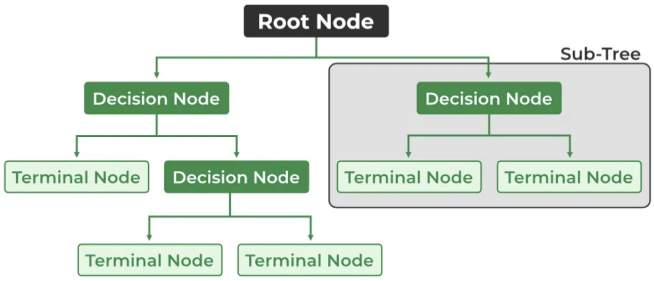


Figure 2- Decision Tree

3. **K-Nearest Neighbors (KNN):** K-Nearest Neighbors is a lazy learning algorithm used for classification and regression tasks. KNN predicts the target value based on the majority vote (classification) or averaging (regression) of its 'k' nearest neighbors in the feature space. KNN is simple, effective for small to medium-sized datasets, and works well in cases where data doesn't have a clear boundary. However, it can be computationally expensive for large datasets as it requires calculating distances for each prediction.

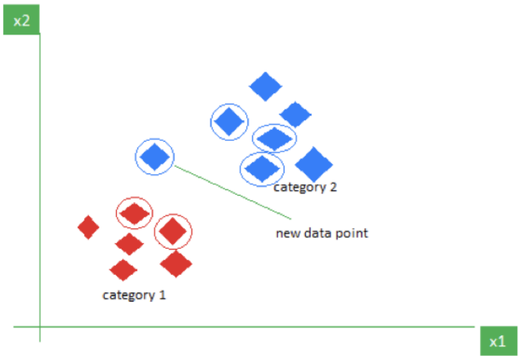


Figure 3 – KNN algorithm working visualization

4. **Linear Regression:** Linear Regression models the relationship between input features and a target variable assuming a linear relationship. It calculates coefficients for each feature to predict continuous outcomes. Linear Regression is interpretable, computationally efficient, and provides insights into feature significance. However, its performance might be limited in capturing intricate nonlinear relationships present in real-world data.

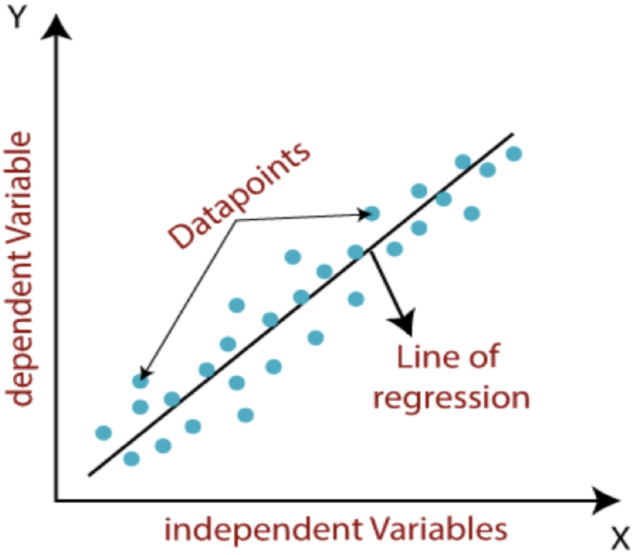


Figure 3- Linear Regression

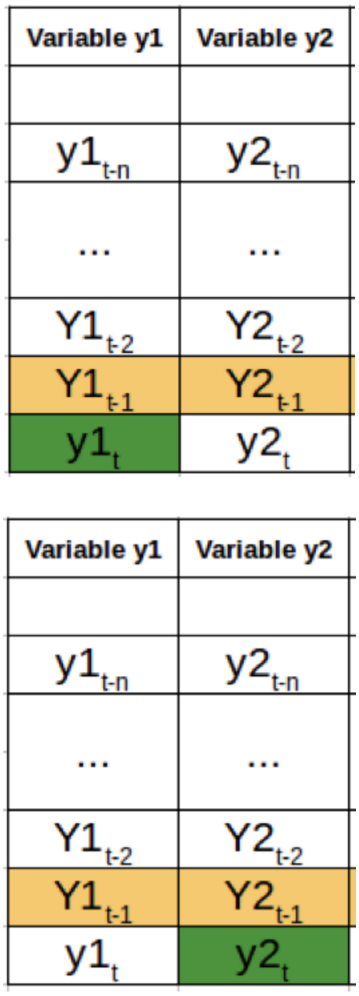
5. **Vector Autoregression (VAR):** VAR models are used to analyze relationships among multiple time series variables. They capture the linear interdependencies between variables by representing each variable as a function of its lagged values and the lagged values of other variables in the system. VAR models are well-suited for understanding the dynamics of multivariate time series data and forecasting each variable's future values based on its own past values and the past values of other variables in the system.

Dealing With a Multivariate Time Series – VAR

8Vector Auto Regression (VAR).

In a VAR algorithm, each variable is a linear function of the past values of itself and the past values of all the other variables. To explain this in a better manner, I'm going to use a simple visual example:

We have two variables, y1, and y2. We need to forecast the value of these two variables at a time 't' from the given data for past n values. For simplicity, I have considered the lag value to be 1.



Variable y1	Variable y2
$y1_{t-n}$	$y2_{t-n}$
…	…
$Y1_{t-2}$	$Y2_{t-2}$
$Y1_{t-1}$	$Y2_{t-1}$
$y1_{t}$	$y2_{t}$

Variable y1	Variable y2
$y1_{t-n}$	$y2_{t-n}$
…	…
$Y1_{t-2}$	$Y2_{t-2}$
$Y1_{t-1}$	$Y2_{t-1}$
$y1_{t}$	$y2_{t}$

Figure 5 – Multivariate Time Series – VAR

Simple mathematical way of representing this relation:

$$y_1(t) = a_1 + w_1 11 * y_1(t-1) + w_1 12 * y_2(t-1) + e_1(t-1)$$

$$y_2(t) = a_2 + w_1 21 * y_1(t-1) + w_1 22 * y_2(t-1) + e_2(t-1)$$

Here,

- a1 and a2 are the constant terms,
- w11, w12, w21, and w22 are the coefficients,
- e1 and e2 are the error terms.

6. **VARIMA (Vector Autoregressive Integrated Moving Average):** VARIMA extends VAR by incorporating differencing to make the time series stationary. By combining autoregression (AR), differencing (I), and moving average (MA) components into a single model, VARIMA can capture complex temporal dependencies, trends, and seasonality within multivariate time series data. This model is especially useful when dealing with non-stationary time series data that exhibit trends or seasonality.

The VARIMA(p, d, q) model is represented as:

For a univariate time series, an ARIMA(p, d, q) model can be denoted as:

ϕp(B)(1−B)dXt=θq(B)Ztϕp (B)(1−B)dXt =θq (B)Zt

Where:

- *XtXt* is the time series at time 't'.
- *ZtZt* is the white noise error term.
- *ϕp(B)ϕp* (*B*)is the autoregressive operator, where *p represents* the order of the autoregressive part.
- *(1−B)d*(1−*B*)*d*is the differencing operator, where *d* represents the order of differencing.
- *θq(B)θq* (*B*)is the moving average operator, where *q* represents the order of the moving average part.
- *BB*is the backshift operator, *BdXt=Xt−dBdXt =Xt−d* .

For the multivariate VARIMA(p, d, q) model, the equations become a system of equations since it's applied to multiple time series variables simultaneously.

The model's equations incorporate lagged values of each variable, differencing, and moving average terms. The VARIMA model is expressed as a system of linear equations involving lagged values of the variables.

7. **Vector Autoregressive Moving-Average (VARMA):** VARMA models extend VAR and VARIMA by combining autoregressive and moving average components. VARMA models capture the dependencies among multiple time series variables by considering both the variables' own lagged values and the lagged forecast errors of other variables in the system. This makes VARMA suitable for modelling multivariate time series data exhibiting both temporal dependencies and residual correlations.

Vector Autoregressive Moving Average (VARMA) model is an extension of the Vector Autoregressive (VAR) model that includes moving average terms. The VARMA(p, q) model is represented as:

Xt=c+Φ1Xt−1+…+ΦpXt−p+Zt+Θ1Zt−1+…+ΘqZt−qXt =*c*+Φ1 *Xt*−1 +…+Φ*p Xt*−*p* +*Zt* +Θ1 *Zt*−1 +…+Θ*q Zt*−*q*

Where:

- *XtXt* is a vector of time series variables at time 't'.
- *cc*is a constant or a vector of constants.
- *Φ1,Φ2,…,Φp*Φ1 ,Φ2 ,…,Φ*p* are coefficient matrices for autoregressive terms up to lag *pp*.
- *ZtZt* is a vector of white noise error terms at time 't'.

- $\Theta 1,\Theta 2,\ldots,\Theta q\Theta 1$,$\Theta 2$,…,Θq are coefficient matrices for moving average terms up to lag qq.

This equation represents a system of equations for each time series variable, where each equation is a function of lagged values of all variables and lagged errors. The model captures the linear dependencies among multiple time series variables by incorporating both autoregressive and moving average components.

8. **Long Short-Term Memory (LSTM):** LSTM, another variant of RNNs, is designed to overcome the limitations of traditional RNNs in capturing long-range dependencies. LSTMs utilize a more complex architecture with memory cells and gates to selectively remember or forget information over long sequences, making them effective in modeling and predicting sequences with long-term dependencies.

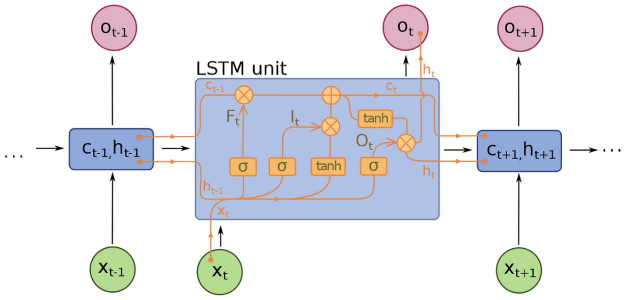


Figure 6 – LSTM Architecture

IV. FRONT END DESIGN ARCHITECTURE

The frontend interface serves as the pivotal link between users and our commodity price prediction system. With an emphasis on user-centric design, it integrates captivating visuals, interactive trend displays, and real-time updates, ensuring an immersive and seamless experience. The Home Page, designed strategically, offers clear navigation through a professional header and engages users with visually appealing hero sections displaying essential project details. Trend graphs facilitate insights into dynamic price trends, while a user-friendly commodity selection feature simplifies choices among 14 commodities. Real-time updates on markets and indices are presented in an organized grid layout for easy consumption.

The Prediction Page provides an interactive platform enabling effective forecasting. Users input commodity data, receive predictions, visualize them alongside historical data, and explore correlations with economic indicators. Moreover, downloadable reports in various formats offer comprehensive insights, empowering informed decision-making in commodity trading and investments.

The About Us Page introduces the project and the team behind it. Detailed team profiles, project overview, and contact information foster familiarity and credibility. Adhering to design guidelines, the CSS implementation ensures consistency across sections, promoting professionalism and ease of use.

Technical implementation involves translating design concepts into live systems using HTML, CSS, and JavaScript. Emphasizing user experience, the frontend design prioritizes usability, accessibility, and responsiveness across devices.

In conclusion, the frontend's successful implementation, aligning with design guidelines and user experience considerations, significantly enhances user engagement and system usability. Future enhancements are envisaged to further elevate user interaction and satisfaction.

VI. FRONT END ARCHITECTURAL DIAGRAM OVERVIEW:

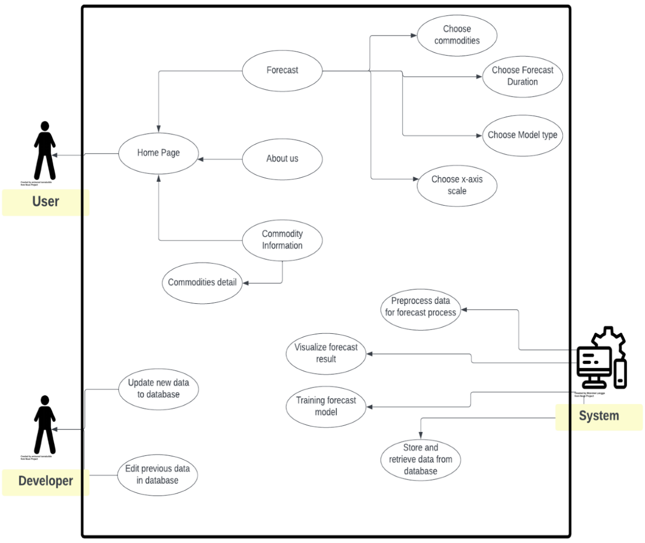


Figure 7 – Frontend Architectural Diagram

VII. RESULTS AND DISCUSSION

The Results and Discussions section meticulously evaluates diverse predictive models employed in forecasting commodity prices. Delving into critical performance metrics including Mean Squared Error

(MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared, it rigorously assesses the efficacy and reliability of each model. These metrics collectively paint a comprehensive picture of prediction accuracy, model fit, and explanatory power.

The models under scrutiny encompass a spectrum of approaches: Support Vector Machine (SVM), Decision Tree Regressor, K Neighbors Regressor, Linear Regression, Vector Autoregression (VAR), VARIMA, VARMA, and Long Short-Term Memory (LSTM). The discussion dissects their performance, strengths, weaknesses, and practical implications, aiming to pinpoint optimal models for real-world applications within financial markets.

The evaluation metrics, such as MSE, RMSE, MAE, and R-squared, serve as crucial indicators. Mean Squared Error quantifies the average squared difference between predicted and actual values, while RMSE, its square root, offers a more interpretable metric in the same units as the target variable. MAE, by calculating the average of absolute differences between predicted and actual values, offers a more straightforward interpretation of accuracy. R-squared measures how well the model explains the variance in the target variable, providing insights into model fit.

The comparative analysis showcases varying model performances: the Decision Tree Regressor demonstrates exceptional accuracy, elucidating nearly 99% of the variance in commodity prices, while the Linear Regression model achieves near-perfect performance, potentially indicating overfitting. SVM and LSTM models display moderate errors and good explanatory power, while the K Neighbors Regressor and VAR, VARIMA, VARMA models exhibit higher errors and relatively poorer fits.

In-depth exploration of these models aids in identifying suitable models for accurate and reliable commodity price prediction. Stakeholders and researchers benefit from this comprehensive analysis, as it pinpoints the strengths and weaknesses of each model, offering insights crucial for practical applications within financial markets. The meticulous evaluation, encompassing performance metrics and model implications, presents a robust foundation for informed decision-making in predictive modeling within the financial domain.

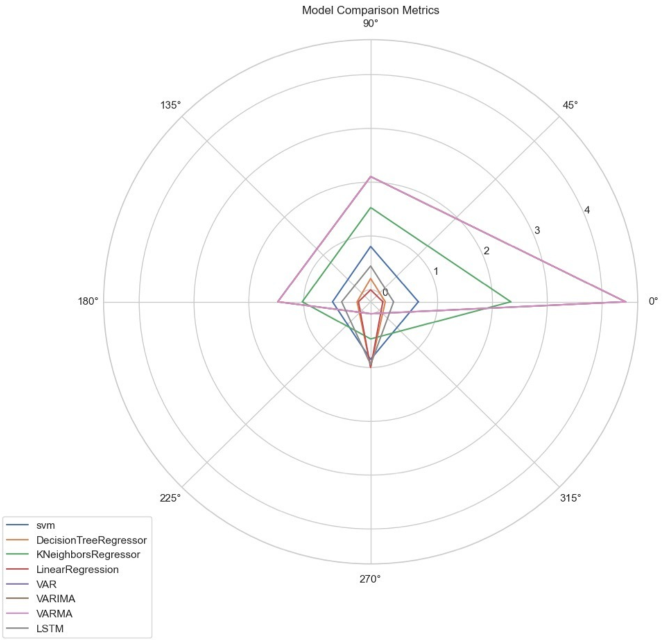


Figure 8

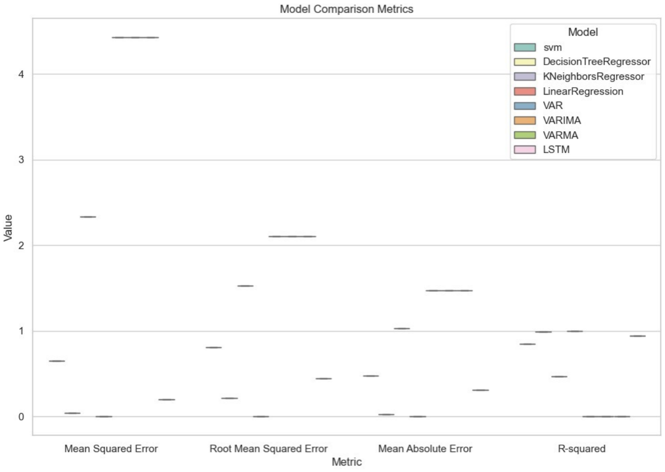


Figure 9

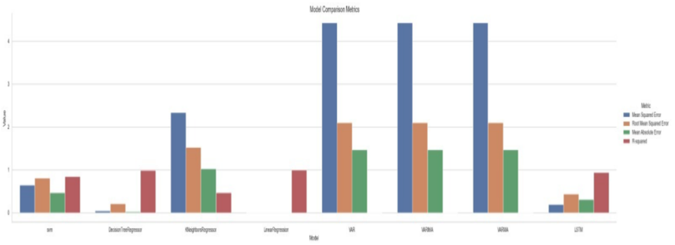


Figure 10

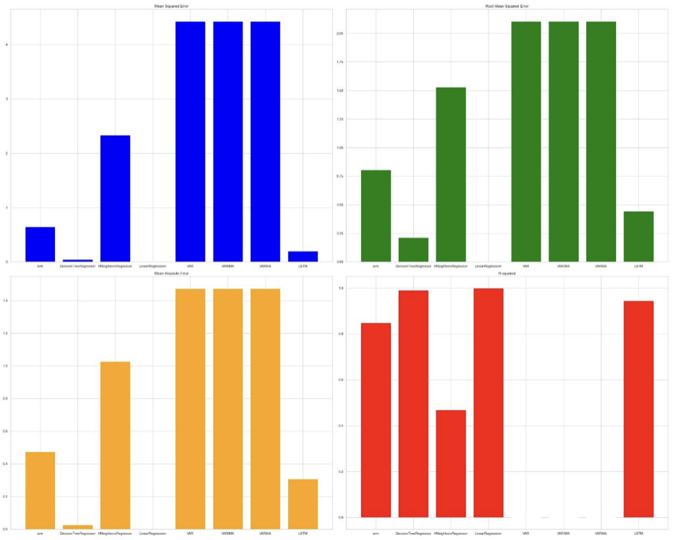


Figure 11

VIII. CONCLUSION

In conclusion, this project has traversed the complex terrain of commodity price prediction, employing a multifaceted strategy encompassing data collection, preprocessing, modeling, and frontend design. This journey underscored the crucial role of accurate predictions within the volatile commodity market landscape. By integrating diverse economic indicators and employing a spectrum of machine learning models, the aim was to unravel the enigmatic nature of commodity price movements.

This venture highlighted the pivotal significance of data in empowering predictive models. The relentless pursuit of high-quality, diverse datasets and meticulous preprocessing unveiled the profound impact of data quality on model efficacy. Feature engineering, combined with a deep understanding of market dynamics, breathed life into the models, enabling them to forecast with considerable accuracy.

The evaluation of models formed an integral part of this expedition, shedding light on their strengths and limitations. Comparative analyses of SVM, Decision Trees, LSTM, and other models underscored the importance of selecting the right model aligned with the dataset's inherent characteristics. The outcomes gleaned from this journey transcend mere predictions, symbolizing a shift towards informed decision-making within financial markets. The intricately designed and calibrated interface stands as evidence of merging predictive capabilities with user-friendly interactions, empowering stakeholders to navigate market turbulences with insightful perspectives.

While this project culminates here, its impact resonates beyond these findings. The outlined recommendations pave the way for future research, probing unexplored facets of commodity markets, embracing advanced modeling techniques, and refining data collection methodologies. The comprehensive evaluation of diverse models for commodity price prediction illuminates distinctive performances across multiple metrics. Among the assessed models, the Decision Tree Regressor and LSTM models emerge as standouts, showcasing exceptional predictive capabilities. Their adeptness at discerning intricate patterns within commodity price data underscores their potential as reliable forecasting tools amid market volatility.

However, cautious interpretation of results remains crucial. While Linear Regression demonstrates near-perfect alignment with the dataset, reflected in its perfect R-squared value, concerns about potential overfitting arise. This necessitates a delicate balance between model accuracy and robustness for reliable real-world predictions. Conversely, the K Neighbors Regressor and VAR, VARIMA, VARMA models exhibit relatively weaker performances, highlighting the challenges in accurately predicting commodity prices through these methodologies.

This nuanced analysis emphasizes the pivotal role of model selection and comprehension of performance metrics in commodity price forecasting. The robust accuracies of the Decision Tree Regressor and LSTM models offer promising avenues for stakeholders seeking reliable insights amid market volatility. However, continual exploration and refinement of alternative approaches remain imperative to further enhance predictive capabilities. In essence, this comprehensive evaluation not only sheds light on the nuanced performances of various models but also underscores the significance of informed assessments. These assessments are crucial in guiding stakeholders towards astute decisions within the dynamic landscape of commodity markets, where precision and reliability are paramount.

III. REFERENCES

[1] Chen, Y., & So, M. K. P. (2020). Machine learning in commodity markets. Applied Economics, 52(56), 6139-6160.
[2] Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. Neurocomputing, 10(3), 215-236.
[3] Lütkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Springer Science & Business Media.
[4] Mohanty, R., Kougianos, E., Yang, C., & Jha, R. K. (2016). Multi-Sensor Data Fusion Using Deep Learning: A Survey. IEEE Access, 4, 3491-3508.
[5] Salisu, A. A., & Raheem, I. D. (2018). Modelling oil price–US stock nexus: A VARMA–BEKK–AGARCH approach. Energy Economics, 76, 1-19.

[6] Wang, D., Smith, K. A., & Hyndman, R. J. (2006). Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery, 13(3), 335-364.

[7] Yang, Y., & Tan, Y. (2019). Commodity Prices Prediction Based on Machine Learning Algorithms. Journal of Physics: Conference Series, 1261(1), 012028.

[8] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

[9] Zou, Y., & Wang, Y. (2019). Forecasting the Prices of Commodities Based on SVR Model. In Proceedings of the 2019 International Conference on Management, Education Technology and Economics (METE 2019) (pp. 108-111). Atlantis Press.

[10] Zurada, J. M. (1992). Introduction to artificial neural systems. West Publishing Company.

[11] Analytics Vidhya. (n.d.). Online platform for learning and competing in data science. Retrieved from [Analytics Vidhya website]: https://www.analyticsvidhya.com/