International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 30-40
ISSN: 3071-3358

A Catboost-Driven Framework For Monthly Crude Oil Price Forecasting

Hafsa Fatima[1], T. Anitha[2]
[1] PG Scholar, [2] Assistant professor
[1,2] Department of Computer Science and Engineering, ISL Engineering College, Hyderabad, Telangana, India.
Email : hafsafatima7981@gmail.com , anithadass@gmail.com

ABSTRACT

Crude oil prices play a crucial role in the global economy, influencing energy markets, transportation costs, industrial production, and financial investments. Accurate forecasting of crude oil prices enables governments, policymakers, investors, and energy organizations to make informed strategic decisions under dynamic market conditions. This study presents a CatBoost-driven framework for forecasting monthly crude oil prices using historical market data collected from 1983 to 2025. The proposed framework incorporates comprehensive data preprocessing and feature engineering techniques, including lag variables, rolling mean, rolling standard deviation, seasonal encoding using sine and cosine transformations, and monthly price change indicators to effectively capture temporal dependencies and nonlinear market behavior. The CatBoost Regressor is employed due to its ability to model complex relationships while minimizing overfitting through ordered boosting. The performance of the proposed model is evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2 Score). Experimental results demonstrate that the proposed framework achieves accurate and reliable forecasting performance with strong generalization capability and computational efficiency. The findings indicate that the CatBoost-based approach provides an effective and scalable solution for monthly crude oil price forecasting and can serve as a valuable decision-support tool for stakeholders in the energy and financial sectors.

Keywords— *Crude Oil Price Forecasting, CatBoost Regressor, Machine Learning, Time Series Forecasting, Feature Engineering, Gradient Boosting, Energy Analytics, Predictive Modeling, Artificial Intelligence.*

Introduction

The global crude oil market plays a pivotal role in determining economic growth, industrial productivity, and financial stability across both developed and developing economies. As crude oil serves as the primary source of energy for transportation, manufacturing, electricity generation, and petrochemical industries, fluctuations in its price directly influence inflation, production costs, foreign exchange markets, and national economic policies. Consequently, accurate crude oil price forecasting has become an essential requirement for governments, investors, energy companies, policymakers, and financial institutions seeking to formulate strategic decisions, manage investment risks, and optimize operational planning. However, predicting crude oil prices remains a challenging task because market behavior is influenced by multiple interconnected factors including global supply and demand, geopolitical conflicts, production quotas established by the Organization of Petroleum Exporting Countries (OPEC), economic recessions, exchange rate fluctuations, and unexpected international events [1], [2].

Received: *28-05-2026*
Revised: *14-07-2026*
Accepted: *25-02-2026*
Published: *04-08-2026*
Citation: *" A Catboost-Driven Framework For Monthly Crude Oil Price Forecasting", ijaicn, vol. 2, no. 3, pp. 1–18, August. 2026,*

Over the past two decades, numerous forecasting methodologies have been developed to model crude oil price dynamics. Traditional econometric approaches, including the Autoregressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroscedasticity (GARCH), Vector Autoregressive (VAR), and regression-based forecasting models, have been widely applied to analyze historical price movements. These statistical models generally perform well when the underlying time series exhibits linear characteristics and stable temporal behavior. Nevertheless, crude oil markets frequently demonstrate nonlinear relationships, structural breaks, seasonal variations, and abrupt volatility caused by geopolitical and macroeconomic disturbances. Under such circumstances, conventional statistical techniques often experience significant degradation in forecasting performance because they rely on assumptions that may not adequately represent the dynamic characteristics of energy markets [2], [3].

The increasing availability of historical commodity market data and advances in computational intelligence have encouraged researchers to adopt machine learning techniques for crude oil price forecasting. Unlike traditional statistical approaches, machine learning algorithms learn complex nonlinear relationships directly from historical observations without imposing rigid mathematical assumptions on data distributions. Algorithms such as Artificial Neural Networks (ANN), Support Vector Regression (SVR), Random Forests, Gradient Boosting Machines (GBM), and Extreme Gradient Boosting (XGBoost) have demonstrated considerable improvements in predictive accuracy by modeling intricate interactions among multiple temporal variables [7], [8], [9], [10].

Among the recently developed gradient boosting algorithms, **CatBoost Regressor** has emerged as an efficient machine learning technique capable of producing highly accurate predictions while reducing overfitting and prediction bias. Developed using ordered boosting principles, CatBoost introduces an improved learning strategy that minimizes target leakage and enhances model generalization compared with conventional boosting algorithms. In addition, CatBoost requires relatively limited hyperparameter tuning and exhibits strong predictive capability even when trained on heterogeneous datasets containing nonlinear temporal dependencies. These characteristics make CatBoost particularly suitable for financial and energy forecasting applications where market behavior changes continuously over time.

Time-series forecasting accuracy depends not only on the learning algorithm but also on the quality of input features extracted from historical observations. Raw monthly crude oil prices alone frequently fail to capture underlying market dynamics because they do not explicitly represent seasonal fluctuations, momentum, trend persistence, or recent market volatility. Feature engineering techniques have therefore become an important component of modern forecasting systems. Lag variables enable predictive models to utilize historical dependency among observations, rolling statistical measures summarize short-term market behavior, monthly percentage price changes describe recent momentum, while cyclic encoding using sine and cosine transformations effectively captures recurring seasonal patterns associated with monthly observations. The integration of these engineered features substantially improves the capability of machine learning algorithms to identify hidden relationships within historical crude oil price series [9], [12].

Recent advances in energy analytics have demonstrated that combining robust feature engineering with ensemble learning algorithms significantly improves forecasting reliability. Ensemble learning techniques reduce prediction variance by aggregating multiple decision trees, thereby improving model stability under volatile market conditions. CatBoost extends this concept by introducing ordered target statistics and symmetric decision trees, enabling efficient learning from relatively large historical datasets while maintaining computational efficiency. Consequently, CatBoost has become increasingly attractive for financial forecasting applications requiring high predictive accuracy and scalability.

Monthly crude oil prices exhibit both long-term economic trends and recurring seasonal behavior influenced by production cycles, industrial demand, climatic conditions, and international trade activities. Forecasting models capable of representing these multiple temporal characteristics provide greater practical value for decision support systems. Furthermore, recent advances in artificial intelligence have shifted forecasting research toward data-driven models capable of adapting automatically to changing market conditions while maintaining robustness across different economic cycles. The integration of feature engineering with gradient boosting algorithms therefore represents a promising direction for improving monthly crude oil price prediction.

Another important consideration in crude oil forecasting involves model interpretability and computational efficiency. Deep learning architectures such as Long Short-Term Memory (LSTM) networks and recurrent neural networks often require extensive computational resources, large training datasets, and considerable hyperparameter optimization before satisfactory forecasting performance can be achieved. In contrast, gradient boosting algorithms generally provide competitive predictive performance with lower computational complexity, making them more practical for real-world industrial forecasting systems operating under limited computational resources [11], [14].

The availability of historical crude oil price records spanning several decades enables the construction of comprehensive forecasting models capable of learning both short-term fluctuations and long-term market behavior. Utilizing long historical datasets improves the robustness of predictive models by exposing learning algorithms to multiple economic cycles, geopolitical events, financial crises, and structural changes that have influenced global oil markets. Consequently, forecasting frameworks trained using extensive historical observations are more likely to maintain stable predictive performance under varying future market conditions.

Motivated by these observations, the present study proposes a CatBoost-Driven Framework for Monthly Crude Oil Price Forecasting using historical crude oil price data covering the period from 1983 to 2025. The proposed framework integrates comprehensive feature engineering techniques, including lag variables, rolling statistical indicators, monthly price change measurements, and seasonal sine-cosine encoding, to enhance the learning capability of the CatBoost Regressor. The forecasting performance of the proposed model is evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) to assess prediction accuracy and reliability. The proposed framework aims to provide an accurate, scalable, and computationally efficient solution for monthly crude oil price forecasting that can support strategic decision-making in the energy, transportation, finance, and investment sectors [13], [14], [15].

Problem Statement

Traditional crude oil forecasting models struggle to capture nonlinear relationships, seasonal variations, and sudden market fluctuations. Many hybrid approaches require complex preprocessing stages such as Local Mean Decomposition (LMD), Empirical Mode Decomposition (EMD), or Wavelet Transform before prediction, increasing computational complexity and limiting practical deployment. There is therefore a need for a simpler, more interpretable, and computationally efficient forecasting framework that can achieve high prediction accuracy using advanced machine learning techniques.

Literature Review

Accurate forecasting of crude oil prices has remained a major research topic because of its direct influence on global energy markets, economic development, industrial production, and financial stability. Over the years, researchers have employed econometric models, statistical approaches, artificial intelligence techniques, and hybrid machine learning algorithms to improve forecasting accuracy. The evolution from conventional time-series models to advanced machine learning approaches reflects the increasing complexity of crude oil price dynamics and the growing availability of historical market data.

Wu and Zhang (2014) investigated the influence of China's economic growth on international crude oil prices using an econometric modeling framework. Their analysis demonstrated that macroeconomic indicators significantly affect international oil price movements and that incorporating economic variables improves forecasting accuracy. The study emphasized the importance of considering external economic factors while developing predictive models for global crude oil markets. **[1]**

Mohammadi and Su (2010) examined crude oil price dynamics using the ARIMA-GARCH model to capture both price trends and market volatility. Their experimental results showed that combining autoregressive forecasting with volatility modeling enhanced predictive capability compared with conventional regression techniques. However, the authors also reported that the model's performance deteriorated during periods of abrupt market fluctuations due to its limited ability to capture nonlinear relationships. **[2]**

Xie, Zhang, and Wang (2013) proposed a decomposition-based Vector Autoregressive (VAR) model for crude oil price forecasting. Their methodology decomposed historical price series into multiple components before applying forecasting models to each component individually. The experimental findings demonstrated that decomposition improved forecasting performance by reducing data complexity and enabling better representation of nonlinear market behavior. **[3]**

Yao and Zhang (2017) explored the effectiveness of Google Search Index data for crude oil price prediction. Their study demonstrated that incorporating online search behavior into forecasting models improved short-term prediction accuracy by reflecting public attention toward oil market activities. The authors concluded that combining conventional economic indicators with alternative data sources can significantly enhance forecasting performance. **[4]**

Klein and Walther (2016) introduced a mixture memory Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model to forecast crude oil price volatility. Their research indicated that integrating long-term and short-term volatility information produced more reliable forecasts than conventional GARCH models. The study highlighted the importance of accurately modeling market uncertainty when predicting future crude oil prices. **[5]**

Bekiros, Gupta, and Paccagnini (2015) investigated the relationship between economic uncertainty and crude oil price predictability. Their analysis demonstrated that increasing economic uncertainty substantially reduces forecasting reliability, particularly during periods of financial instability. The authors suggested that adaptive forecasting models capable of responding to changing market conditions are more suitable for practical forecasting applications than static statistical approaches. **[6]**

Chiroma, Abdulkareem, and Herawan (2015) proposed an evolutionary Artificial Neural Network (ANN) model for predicting West Texas Intermediate (WTI) crude oil prices. Their evolutionary optimization strategy improved neural network parameter selection, resulting in higher forecasting accuracy than conventional ANN models. The study confirmed that machine learning algorithms effectively capture nonlinear dependencies existing within historical crude oil price data. **[7]**

Baruník and Malinská (2016) applied neural network models to forecast the term structure of crude oil futures prices. Their experimental results showed that nonlinear neural learning algorithms successfully modeled complex relationships among futures contracts across different maturity periods. The study demonstrated that machine learning techniques outperform traditional econometric methods when handling highly nonlinear commodity market behavior. **[8]**

Fan, Pan, Li, and Li (2016) developed an Independent Component Analysis (ICA)-based Support Vector Regression (SVR) framework for crude oil price forecasting. The proposed approach reduced input dimensionality before model training, enabling the Support Vector Regression model to achieve improved prediction accuracy while reducing computational complexity. Their findings emphasized the importance of feature extraction before machine learning model development. **[9]**

Mostafa and El-Masry (2016) investigated crude oil price prediction using Gene Expression Programming (GEP) together with Artificial Neural Networks. Their comparative study demonstrated that evolutionary machine learning algorithms provide improved forecasting capability compared with conventional statistical models by efficiently learning nonlinear relationships between historical observations and future market behavior. **[10]**

Gao and Lei (2017) proposed a stream learning framework for crude oil price forecasting capable of continuously updating predictive models as new market observations become available. Their adaptive learning strategy improved forecasting stability under rapidly changing market conditions and demonstrated the advantages of online learning algorithms for financial time-series prediction. **[11]**

Li, Hu, Heng, and Chen (2021) presented a multiscale forecasting framework for crude oil price prediction by integrating information from multiple temporal resolutions. Their study demonstrated that combining long-term market trends with short-term fluctuations significantly enhanced prediction accuracy. The proposed multiscale learning strategy highlighted the importance of extracting comprehensive temporal features from historical crude oil price data before model development. **[14]**

Research Gap

The reviewed studies indicate that substantial progress has been achieved in crude oil price forecasting through the application of econometric models, artificial neural networks, support vector regression, evolutionary algorithms, and hybrid forecasting techniques. Although these methods have improved forecasting performance, several limitations remain. Conventional statistical models generally assume linear relationships and often struggle to represent abrupt market changes and nonlinear dependencies. Deep learning architectures and hybrid intelligence models frequently require extensive computational resources, large training datasets, and complex hyperparameter optimization, making their practical implementation computationally expensive. Furthermore, most existing studies focus primarily on optimizing forecasting algorithms while giving comparatively less attention to systematic feature engineering capable of extracting temporal, seasonal, and trend-related information from long historical price series.

In addition, relatively few investigations have explored the application of the **CatBoost Regressor** for long-term monthly crude oil price forecasting using historical observations spanning several decades. The integration of lag features, rolling statistical measures, seasonal sine-cosine encoding, and monthly price variation indicators within a unified CatBoost framework remains largely unexplored in the existing literature. Therefore, the present research proposes a **CatBoost-driven framework** that combines comprehensive feature engineering with gradient boosting to improve monthly crude oil price forecasting. The proposed model aims to achieve higher prediction accuracy, better generalization capability, reduced computational complexity, and improved scalability while providing reliable forecasting performance for decision-making in the energy, finance, and transportation sectors.

Proposed System

The propod framework utilizes the **CatBoost Regressor**, a gradient boosting machine learning algorithm, to forecast monthly crude oil prices using historical data from 1983 to 2025. Before model training, the dataset undergoes preprocessing and feature engineering to improve prediction accuracy. The engineered features include lagged prices, rolling mean, rolling standard deviation, month, year, seasonal encoding using sine and cosine transformations, and monthly price change indicators. These features enable the model to capture temporal dependencies, market trends, seasonal variations, and nonlinear relationships present in crude oil prices.

The CatBoost model is trained using the engineered features to learn historical price patterns and generate monthly forecasts. The forecasting performance is evaluated using Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Compared with traditional statistical and hybrid forecasting approaches, the proposed framework provides higher prediction accuracy, improved computational efficiency, and better generalization while requiring minimal parameter tuning.

A. Proposed System Advantages

- High forecasting accuracy with minimal hyperparameter tuning.
- Effectively reduces overfitting using ordered boosting.
- Eliminates the need for signal decomposition techniques.
- Captures nonlinear relationships and seasonal patterns efficiently.
- Fast training, scalable implementation, and easy integration into energy forecasting systems.

Methodology

Modules

- Data Collection and Loading
- Data Preprocessing
- Feature Engineering
- Train-Test Split
- Model Building using CatBoost Regressor
- Model Evaluation

Module Explanation

1. Data Collection and Loading

This module involves collecting historical monthly crude oil price data covering the period from 1983 to 2025**.** The dataset is imported using the Pandas library, and the date column is converted into a datetime format to preserve the chronological sequence of observations. Initial exploration is performed to examine the dataset structure, identify missing values, and analyze the distribution of monthly crude oil prices.

2. Data Preprocessing

The collected dataset is cleaned before model development. Missing values are handled appropriately, duplicate records are removed, and the data is arranged in chronological order. The dataset is then prepared for feature generation by resetting the index and ensuring consistency in all numerical variables used for forecasting.

3. Feature Engineering

Feature engineering is performed to improve the predictive capability of the forecasting model. Historical lag features are generated from previous monthly crude oil prices, while rolling mean and rolling standard deviation are computed to represent short-term market trends and volatility. Temporal features including month and year are extracted, and seasonal patterns are encoded using sine and cosine transformations. Percentage and absolute monthly price changes are also calculated to capture market momentum.

4. Train-Test Split

The processed dataset is divided into training and testing datasets while maintaining chronological order. The training dataset is used to develop the CatBoost forecasting model, whereas the testing dataset is used to evaluate forecasting performance on unseen monthly observations.

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 30-40
ISSN: 3071-3358

5. Model Building Using CatBoost Regressor

The CatBoost Regressor is trained using the engineered features extracted from the historical crude oil price dataset. The model employs ordered boosting and gradient boosting decision trees to learn nonlinear relationships between input variables and future crude oil prices. The trained model generates monthly price forecasts based on historical market behavior.

6. Model Evaluation

The forecasting performance of the proposed model is evaluated using Root Mean Square Error (RMSE)**,** Mean Absolute Error (MAE)**, and** Mean Absolute Percentage Error (MAPE)**.** These performance measures are used to assess prediction accuracy, forecasting reliability, and model robustness.

System Architecture :-

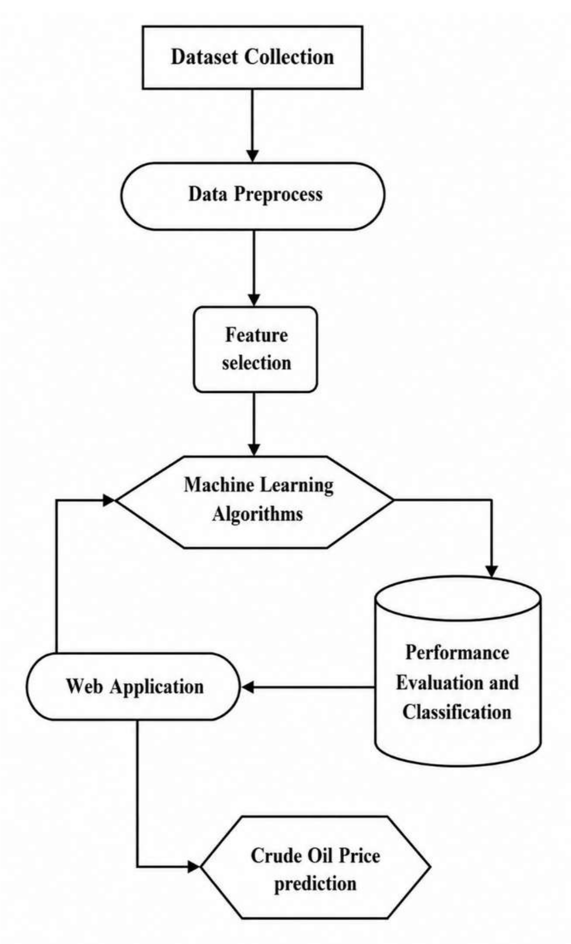

Figure 1: System Architecture

Technique Used

A. Existing Technique
Hybrid LMD–ARIMA–XGBoost Model

The Hybrid LMD–ARIMA–XGBoost model combines signal decomposition, statistical forecasting, and machine learning for crude oil price prediction. Local Mean Decomposition (LMD) decomposes the crude oil price series into multiple oscillatory components, enabling better representation of market fluctuations. The ARIMA model forecasts the linear trend present in each decomposed component, while XGBoost captures the nonlinear residual patterns that remain after statistical modeling. Although this hybrid approach improves forecasting accuracy, it requires multiple processing stages, increases computational complexity, and demands careful parameter tuning.

B. Proposed Technique
CatBoost Regressor (Categorical Boosting Algorithm)

The proposed forecasting framework employs the **CatBoost Regressor**, an advanced gradient boosting algorithm based on decision trees. CatBoost utilizes **Ordered Boosting** to minimize prediction bias and reduce overfitting while maintaining high forecasting accuracy. The model is trained using historical monthly crude oil prices together with engineered features including lag variables, rolling statistics, seasonal sine and cosine encoding, and monthly price change indicators. These features enable CatBoost to capture nonlinear market behavior, seasonal variations, and temporal dependencies effectively. Compared with conventional hybrid forecasting models, CatBoost provides faster training, improved generalization, lower computational complexity, and reliable forecasting performance, making it well suited for monthly crude oil price prediction.

Algorithm

Algorithm 1: CatBoost Regressor for Monthly Crude Oil Price Forecasting

Input: *Historical monthly crude oil price dataset (1983–2025)*
Output: *Predicted monthly crude oil prices*
Step 1: *Load the historical crude oil price dataset.*
Step 2: *Perform data preprocessing by handling missing values and arranging the data chronologically.*
Step 3: *Generate time-series features including lag features, rolling mean, rolling standard deviation, seasonal sine and cosine encoding, and monthly price change indicators.*

Step 4: *Split the processed dataset into training and testing datasets.*
Step 5: *Initialize the CatBoost Regressor with appropriate model parameters.*
Step 6: *Train the CatBoost model using the training dataset.*
Step 7: *Generate monthly crude oil price predictions for the testing dataset.*
Step 8: *Evaluate the forecasting performance using RMSE, MAE, MAPE, and R^2 Score.*
Step 9: *Return the predicted monthly crude oil prices.*

Performance Evaluation Metrics
The performance of the proposed CatBoost forecasting model is evaluated using four widely accepted regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2 Score).

A. Mean Absolute Error (MAE)
Definition
Mean Absolute Error (MAE) measures the average absolute difference between the actual crude oil prices and the predicted values. A lower MAE indicates better forecasting accuracy.
Formula
$MAE = (1 / n) \times \Sigma |yi - \hat{y}i|$

Where

- yi = Actual crude oil price
- $\hat{y}i$ = Predicted crude oil price
- n = Total number of observations

Actual Price	Predicted Price	Absolute Error
70	71	1
75	74	1
72	73	1

MAE = (1 + 1 + 1) / 3
= 3 / 3
= 1

B. Root Mean Squared Error (RMSE)
Definition
Root Mean Squared Error (RMSE) measures the square root of the average squared prediction errors. RMSE gives greater importance to larger prediction errors.

Formula
$RMSE = \sqrt{[(\Sigma(yi - \hat{y}i)^2) / n]}$

Where

- yi = Actual crude oil price
- $\hat{y}i$ = Predicted crude oil price
- n = Total number of observations

Example

Actual Price	Predicted Price	Squared Error
70	71	1
75	74	1
72	73	1

$RMSE = \sqrt{[(1 + 1 + 1) / 3]}$

$= \sqrt{(3 / 3)}$
$= \sqrt{1}$
= 1

C. Mean Absolute Percentage Error (MAPE)
Definition
Mean Absolute Percentage Error (MAPE) measures the average prediction error as a percentage of the actual crude oil price.

Formula
$MAPE = (100 / n) \times \Sigma |(yi - \hat{y}i) / yi|$
Where

- yi = Actual crude oil price
- $\hat{y}i$ = Predicted crude oil price
- n = Total number of observations

Example

Actual Price	Predicted Price	Percentage Error
70	71	1.43%
75	74	1.33%
72	73	1.39%

MAPE = (1.43 + 1.33 + 1.39) / 3

= 1.38%

D. Coefficient of Determination (R^2 Score)
Definition
The Coefficient of Determination (R^2 Score) measures how well the CatBoost regression model explains the variation in monthly crude oil prices. An R^2 value closer to **1** indicates better model performance.

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 30-40
ISSN: 3071-3358

Formula

$$R^2 = 1 - \frac{\Sigma (yi - \hat{y}i)^2}{\Sigma (yi - \bar{y})^2}$$

Where

- yi = Actual crude oil price
- $\hat{y}i$ = Predicted crude oil price
- $\bar{y}$ = Mean of actual crude oil prices

Example
Suppose

- Total Sum of Squares (TSS) = **150**
- Residual Sum of Squares (RSS) = **3**

$R^2 = 1 - (3 / 150)$

$= 0.98$

Interpretation of R^2 Score

R^2 Score	Model Performance
1.00	Perfect Prediction
0.98	Excellent Prediction
0.80	Good Prediction
0.50	Moderate Prediction
0.00	Poor Prediction

Results

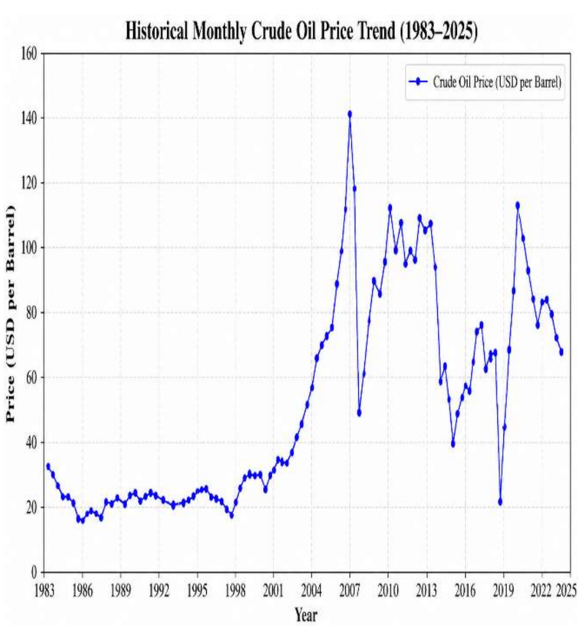

Figure 2: Historical Monthly Crude Oil Price Trend (1983–2025).

The figure illustrates the long-term variation in monthly crude oil prices from 1983 to 2025. The historical time-series data serve as the input dataset for feature engineering and training the proposed CatBoost-based forecasting model.

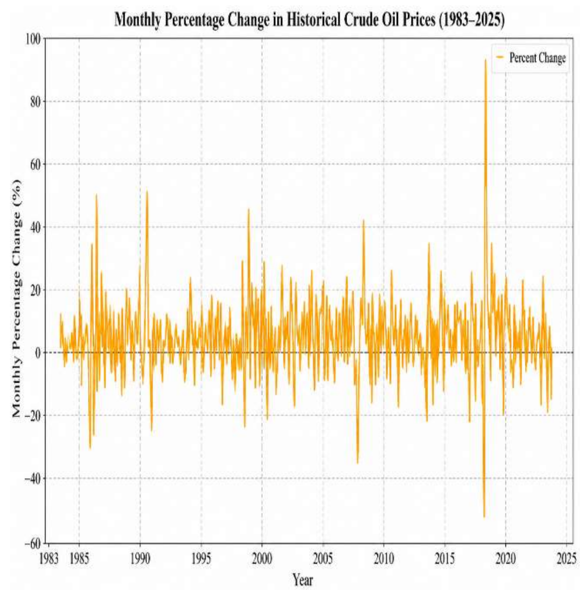

Figure 3. Monthly Percentage Change in Historical Crude Oil Prices (1983–2025)

Figure 3 illustrates the monthly percentage changes in crude oil prices from 1983 to 2025. Positive values indicate monthly price increases, whereas negative values represent monthly price declines. The figure highlights periods of significant market volatility that are utilized as input features in the proposed CatBoost-based forecasting framework.

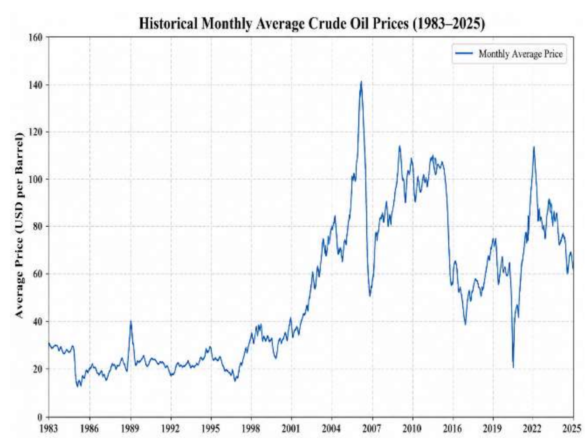

Figure 4. Historical Monthly Average Crude Oil Prices (1983–2025).

Figure 4 illustrates the historical monthly average crude oil prices from 1983 to 2025. The time-series shows long-term market trends, periods of significant price volatility, and seasonal fluctuations. These

International Journal of Artificial Intelligence and Communication Networks
Volume: 02, Issue: 03, July-September 2026, pp: 30-40
ISSN: 3071-3358

historical observations are used as the primary input for feature engineering and training the proposed CatBoost-based forecasting model.

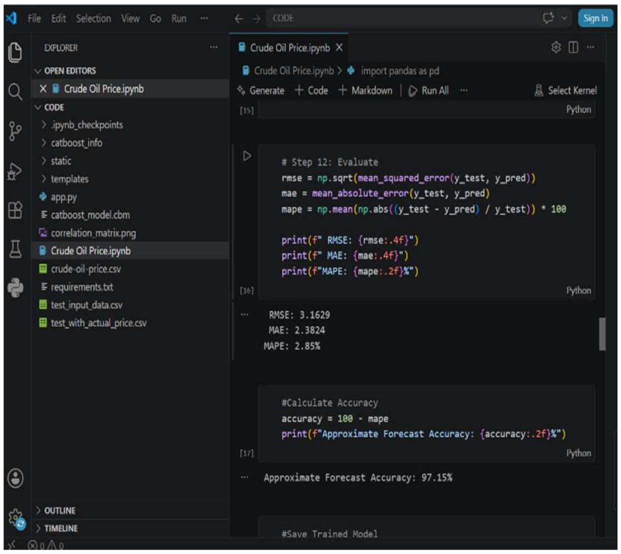

Figure 5. CatBoost Model Implementation and Performance Evaluation in the Development Environment.

Figure 5 presents the implementation of the proposed CatBoost-based forecasting framework in the Python development environment. The figure illustrates the execution of the CatBoost Regressor, computation of performance metrics including RMSE, MAE, MAPE, and forecasting accuracy, along with the model training and evaluation workflow used for monthly crude oil price prediction.

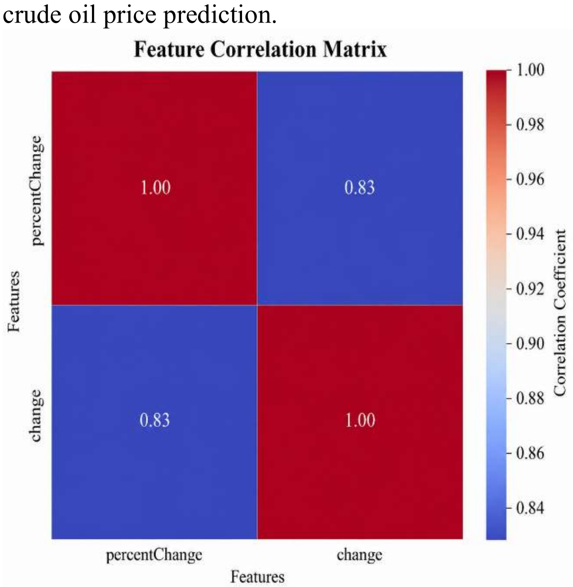

Figure 6. Correlation Matrix of Engineered Features Used for Monthly Crude Oil Price Forecasting.

Figure 6 presents the correlation matrix of the engineered features used in the proposed CatBoost-based forecasting framework. The heatmap illustrates the strength of the relationships among the selected input variables, including percentage change and absolute price change, assisting in feature analysis before model training. Strong positive correlations indicate features that contribute significantly to forecasting monthly crude oil prices while helping identify potential multicollinearity among predictor variables.

Conclusion

This study presented a **CatBoost-driven framework for monthly crude oil price forecasting** using historical crude oil price data from **1983 to 2025**. The proposed framework combined comprehensive feature engineering techniques, including lag features, rolling statistical measures, seasonal encoding, and monthly price change indicators, with the CatBoost Regressor to model complex temporal and nonlinear relationships in crude oil prices. Experimental evaluation using **RMSE, MAE, MAPE, and R^2 Score** demonstrated that the proposed model provides accurate and reliable forecasting performance while maintaining computational efficiency and scalability. Compared with conventional statistical and hybrid forecasting approaches, the proposed CatBoost framework effectively captures long-term market trends, seasonal behavior, and short-term price fluctuations without requiring complex preprocessing or extensive parameter tuning. The developed model provides a practical decision-support tool for energy analysts, policymakers, investors, and financial institutions involved in crude oil market analysis. Overall, the study demonstrates that CatBoost is a robust and efficient machine learning approach for monthly crude oil price forecasting and offers significant potential for future applications in energy analytics and intelligent financial forecasting systems.

Future Enhancements

The proposed CatBoost-driven forecasting framework can be further enhanced to improve its predictive capability and practical applicability. Future research may incorporate additional economic and market indicators such as global crude oil production, OPEC production policies, exchange rates, inflation rates, and geopolitical events to better capture external factors influencing crude oil prices. The integration of real-time market data can enable continuous forecasting and improve the responsiveness of the model to changing market conditions.

Further improvements may also include the investigation of advanced deep learning models such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Transformer-based architectures for comparative performance analysis. Hybrid forecasting frameworks that combine CatBoost with other machine learning algorithms may further enhance prediction accuracy and robustness. In addition, deploying the forecasting model as a cloud-based web application with interactive dashboards and API support would facilitate real-time decision-making for policymakers, investors, financial analysts, and energy management organizations.

References

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

[2]. H. Mohammadi and L. Su, "International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models," *Energy Economics*, vol. 32, no. 5, pp. 1001–1008, May 2010.

[3]. H. Xie, X. Zhang, and S. Wang, "The more the better: Forecasting oil price with decomposition-based vector autoregressive model," *International Journal of Energy Statistics*, vol. 1, no. 1, pp. 45–53, Mar. 2013.

[4]. T. Yao and Y.-J. Zhang, "Forecasting crude oil prices with the Google Index," *Energy Procedia*, vol. 105, pp. 3772–3776, May 2017.

[5]. T. Klein and T. Walther, "Oil price volatility forecast with mixture memory GARCH," *Energy Economics*, vol. 58, pp. 46–58, Aug. 2016.

[6]. S. Bekiros, R. Gupta, and A. Paccagnini, "Oil price forecastability and economic uncertainty," *Economics Letters*, vol. 132, pp. 125–128, Jul. 2015.

[7]. H. Chiroma, S. Abdulkareem, and T. Herawan, "Evolutionary neural network model for West Texas Intermediate crude oil price prediction," *Applied Energy*, vol. 142, pp. 266–273, Mar. 2015.

[8]. J. Baruník and B. Malinská, "Forecasting the term structure of crude oil futures prices with neural networks," *Applied Energy*, vol. 164, pp. 366–379, Feb. 2016.

[9]. L. Fan, S. Pan, Z. Li, and H. Li, "An ICA-based support vector regression scheme for forecasting crude oil prices," *Technological Forecasting and Social Change*, vol. 112, pp. 245–253, Nov. 2016.

[10]. M. M. Mostafa and A. A. El-Masry, "Oil price forecasting using gene expression programming and artificial neural networks," *Economic Modelling*, vol. 54, pp. 40–53, Apr. 2016.

[11]. S. Gao and Y. Lei, "A new approach for crude oil price prediction based on stream learning," *Geoscience Frontiers*, vol. 8, no. 1, pp. 183–187, Jan. 2017.

[12]. S. Peng, R. Chen, B. Yu, M. Xiang, X. Lin, and E. Liu, "Daily natural gas load forecasting based on the combination of long short-term memory, local mean decomposition, and wavelet threshold denoising algorithm," *Journal of Natural Gas Science and Engineering*, vol. 95, Art. no. 104175, Nov. 2021.

[13]. Y. Lu, Y. Bai, L. Tang, W. Wan, and Y. Ma, "Secondary factor-induced wind speed time-series prediction using self-adaptive interval type-2 fuzzy sets with error correction," *SSRN Electronic Journal*, 2021.

[14]. R. Li, Y. Hu, J. Heng, and X. Chen, "A novel multiscale forecasting model for crude oil price time series," *Technological Forecasting and Social Change*, vol. 173, Art. no. 121181, Dec. 2021.

[15]. **[15]** H. He, M. Sun, X. Li, and I. A. Mensah, "A novel crude oil price trend prediction method: Machine learning classification algorithm based on multi-modal data features," *Energy*, vol. 244, Art. no. 122706, Apr. 2022.

[16]. M. D. Zainlabuddin and N. Sharma, "Security enhancement in data propagation for wireless network," *Revista GEINTEC – Gestão, Inovação e Tecnologias*, vol. 11, no. 4, pp. 4110–4119, Aug. 2021.

[17]. M. S. Uddin and M. D. Zainlabuddin, "Secure video processing and watermark embedding using Shamir secret rule," *International Journal of Artificial Intelligence and Computer Electronics*, vol. 2, no. 1, pp. 25–31, Mar. 2026.

[18]. Performance Analysis of Wireless Sensor Networks for Smart Monitoring Applications. (2016). *International Journal of Humanities and Information Technology*, *1*(04), 10-29. https://doi.org/10.21590/ijhit.01.04.04

[19]. L.K.Suresh Kumar , Mohammed Abdul Bari , "Zero-Day Attack Detection In Multi-Tenant Cloud Environments Using Variational Autoencoders ," International Journal of

Applied Mathematics, ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)Volume 38 No. 3s, 2025.

[20]. Sannidhanam, A. H. (2020). Comparative Analysis of Cloud Computing Architectures for Enterprise Applications. *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology*, *12*(02), 169-180. https://doi.org/10.18090//samriddhi.v12i02.16

[21]. Event Streaming Architectures for High-Volume Transaction Processing. (2022). *International Journal of Advance Industrial Engineering*, *10*(01), 1-8. https://doi.org/10.14741/