International Journal of Multidisciplinary Engineering in Current Research – IJMEC

Volume 11, Issue 4s, April-2026, http://ijmec.com/, ISSN: 2456-4265

Full Length Article

A Machine Learning Framework for Monthly Crude Oil Price Prediction with Cat Boost

Syed Mohd Faizan[1], Mohammed Kasadi[2], Dr. Abdul Ahad Afroz[3]

[1,2]B.E.Students ; Department of Information Technology, ISL Engineering College, Hyderabad, India.

[3]Associate Professor; Department of Information Technology, ISL Engineering College, Hyderabad, India.

Mail Id;syedfai363@gmail.com,mohammedkasadi91@gmail.com

Accepted 24-04-2026

Abstract

Crude oil is a globally significant energy resource whose price fluctuations have far-reaching economic and industrial impacts. Accurate forecasting of crude oil prices is crucial for strategic decision-making in sectors such as finance, energy, and transportation.

This project presents a machine learning-based approach to predict monthly crude oil prices using historical market data and engineered time-series features. The model is developed using the CatBoost Regressor, a high-performance gradient boosting algorithm known for its efficiency, accuracy, and ability to handle complex non-linear data.

The predictive features include: Lagged prices from previous months Rolling statistical indicators (mean and standard deviation) Temporal features such as month and year (encoded using sine and cosine transformations to preserve seasonality) Both percentage and absolute monthly price changes, The dataset spans over four decades (1983–2025), ensuring that the model captures long-term patterns and short-term fluctuations. Model performance is evaluated using RMSE, MAE, and MAPE metrics, demonstrating strong predictive accuracy and generalization capability.

Keywords:

Crude Oil Price Prediction, Machine Learning, CatBoost Regressor, Time Series Forecasting, Energy Economics, Predictive Analytics, Gradient Boosting, Historical Market Data, Feature Engineering, Rolling Statistics, Lag Features, Seasonal Analysis, RMSE, MAE, MAPE, Financial Forecasting, Oil Market Analysis, Data Science, Artificial Intelligence, Economic Prediction.

Introduction

Crude oil is one of the most vital commodities in the global economy, serving as a key input for energy production, transportation, manufacturing, and various other sectors. Due to its strategic importance, the price of crude oil is subject to high volatility, influenced by a wide range of factors including geopolitical tensions, supply-demand imbalances, environmental regulations, and macroeconomic trends.

Traditional statistical forecasting models, such as ARIMA and exponential smoothing, have been widely used for time series prediction tasks. However, these models often fall short in capturing the complex, non-linear patterns and seasonal behaviors inherent in commodity markets like crude oil. With advancements in machine learning, more powerful and flexible models have emerged, offering improved predictive performance by learning from historical trends and patterns.

This project proposes a machine learning approach using the CatBoost Regressor, a robust gradient boosting algorithm specifically designed to handle structured data with high accuracy and low overfitting. The model is trained on a long-term monthly crude oil price dataset, incorporating several engineered features such as lagged prices, rolling averages, standard deviation, month and year encoding, and both absolute and percentage price changes.

Scope of the Work

The scope of this project is centered around the development of an accurate and efficient machine learning model to forecast monthly crude oil prices using historical data. The project utilizes the CatBoost Regressor algorithm, which is capable of capturing non-linear patterns and seasonal trends commonly found in time-series data.

The project covers the entire pipeline — from data preprocessing and feature engineering to model training, evaluation, and prediction. The outcome is a

https://doi.org/10.63665/IJMEC.1104s.42
ISSN: 2456-4265
IJMEC 2026

scalable and interpretable forecasting system that can be adapted or extended to other economic indicators or commodities in future work.

Objective

The primary objective of this project is to develop a reliable and accurate machine learning model to forecast monthly crude oil prices using historical data. By leveraging the CatBoost Regressor algorithm and effective time-series feature engineering, the goal is to: Capture complex trends, seasonality, and non-linear behaviors in crude oil price movements

Minimize prediction errors while remaining computationally efficient and interpretable

Support decision-making in sectors such as energy, finance, transportation, and policy planning

Demonstrate that high forecasting accuracy can be achieved without relying on overly complex hybrid or deep learning methods

Existing System & Disadvantages

Traditional crude oil price forecasting systems commonly use ARIMA (AutoRegressive Integrated Moving Average) combined with decomposition methods like EMD or LMD, and enhanced with XGBoost for non-linear components. While effective in certain scenarios, these hybrid systems suffer from:

High computational complexity

Overfitting in complex models

Dependence on decomposition methods

Low interpretability

Inability to capture cyclical features directly

Proposed System & Advantages

The proposed system uses the CatBoost Regressor — a state-of-the-art gradient boosting algorithm developed by Yandex. It eliminates the need for explicit decomposition while still capturing non-linear patterns, trends, and seasonal behaviors effectively. Key advantages include:

High accuracy with minimal tuning

Efficient handling of overfitting via Ordered Boosting

No need for decomposition preprocessing

Captures non-linear and seasonal patterns natively

Fast training and easy integration into production environments

Project Description & Methodology

Modules

The project is structured into six core modules:

Module	Description
1. Data Collection	Import crude oil dataset (1983–2025) using Pandas, perform basic exploration and parse date columns into datetime format.
2. Data Preprocessing	Clean dataset by handling missing values in percentChange and change columns, sort by date, and check for duplicates or anomalies.
3. Feature Engineering	Create lag features, rolling mean/std, sine/cosine encoded month and year, and absolute/percentage monthly price changes.
4. Train-Test Split	Time-based 80/20 split preserving chronological order. Separate feature matrix (X) and target price variable (y).
5. Model Building	Train CatBoost Regressor with tuned hyperparameters (learning rate, iterations). Save trained model for future predictions.
6. Model Evaluation	Assess performance using RMSE, MAE, and MAPE. Validate on manual inputs to confirm practical usability.

Algorithm: CatBoost Regressor

CatBoost (Categorical Boosting) is a state-of-the-art, open-source gradient boosting algorithm based on decision trees, developed by Yandex. It uses Ordered Boosting and Permutation-Driven Techniques to reduce overfitting and improve generalization — key advantages over XGBoost and LightGBM.

For this project, CatBoost is trained using historical crude oil price data along with engineered features such as lag prices, percent changes, rolling statistics, and seasonality transformations (sine and cosine of month). It effectively captures non-linear interactions in the dataset and delivers accurate forecasts with low RMSE and MAPE scores.

System Requirements

Hardware Requirements

Component	Specification

https://doi.org/10.63665/IJMEC.1104s.42
ISSN: 2456-4265
IJMEC 2026

Processor	Intel i5 / i7 or equivalent
RAM	8 GB / 16 GB
Hard Disk	512 GB SSD

Software Requirements

Component	Specification
Operating System	Windows 10/11
Platform / IDE	PyCharm / Jupyter Notebook
Programming Language	Python 3.x
Key Libraries	NumPy, Pandas, Matplotlib, Scikit-learn, CatBoost

System Design

The system design of this project is represented through a series of UML (Unified Modelling Language) diagrams that collectively describe the structure, behavior, and interactions of the crude oil price forecasting system.

System Architecture

The overall system pipeline follows a structured flow:

Dataset Collection — Historical crude oil price data (1983–2025)

Data Preprocessing — Cleaning, sorting, and handling missing values

Feature Selection & Engineering — Lag features, rolling stats, temporal encodings

Machine Learning Algorithm — CatBoost Regressor training

Performance Evaluation — RMSE, MAE, MAPE calculation

Web Application — Flask-based UI for user interaction

Crude Oil Price Prediction — Final output displayed to user

Key UML Diagrams Summary

Use Case Diagram: The system involves a single actor (User) who can register/login, enter input features, trigger data preprocessing and model training, view predicted prices, and review charts and metrics.

Class Diagram: Four main classes are defined — User (authentication), PredictionSystem (model logic), Visualization (chart generation), and Metrics (RMSE/MAE/MAPE calculation). These interact sequentially in the prediction pipeline.

Sequence Diagram: Captures the interaction flow from user login through feature input, CatBoost prediction, and visualization rendering back to the user interface.

Data Flow Diagram (DFD): At Level 0, the user provides inputs and receives prediction results and visualizations. Level 1 breaks down into preprocessing, feature engineering, model prediction, evaluation, and visualization components.

Feature Engineering & Model Details

Engineered Features

Feature engineering is the backbone of this project's accuracy. The following features were derived from the raw monthly price data:

Feature	Description
Lag Features (Lag 1–3)	Price values from 1, 2, and 3 months prior — captures recent momentum and short-term trends
Rolling Mean (3/6 months)	Moving average of recent prices — smooths noise and highlights trend direction
Rolling Std Dev	Standard deviation over rolling window — measures local volatility
Month (sin/cos encoded)	Sine and cosine of month number — preserves cyclical seasonality without ordinal bias
Year	Numeric year — captures long-term secular trends across decades
Absolute Price Change	Month-over-month absolute difference in price

https://doi.org/10.63665/IJMEC.1104s.42
ISSN: 2456-4265
IJMEC 2026

Percentage Price Change	Month-over-month percentage change — normalized metric for market shifts

Model Training & Evaluation

The CatBoost Regressor is trained on 80% of the dataset (chronological split) and tested on the remaining 20%. The model is evaluated using three standard regression metrics:

RMSE (Root Mean Squared Error) — Penalizes large prediction errors; useful for detecting outlier misses

MAE (Mean Absolute Error) — Average magnitude of prediction errors in dollar terms

MAPE (Mean Absolute Percentage Error) — Percentage-based metric enabling comparison across price scales

The results demonstrate strong predictive accuracy and generalization capability, confirming the effectiveness of CatBoost combined with the engineered feature set.

Software Testing

Testing ensures the system meets its functional requirements and performs accurately across all components. The following types of testing were conducted:

Unit Testing

Each module (data loading, preprocessing, feature engineering, model training, evaluation) was tested independently to verify that individual components produce correct outputs given valid inputs.

Functional Testing

Systematic testing confirmed that all specified functions — from user login to prediction display — behave as documented. Both valid and invalid inputs were tested to ensure appropriate acceptance and rejection handling.

Integration Testing

All modules were integrated and tested together to verify seamless data flow from raw input through preprocessing, feature engineering, model prediction, and result visualization without interface errors.

Performance Testing

The system was verified to produce prediction outputs within acceptable time limits. CatBoost's fast training speed (a core design advantage) ensures responsive performance even on standard hardware.

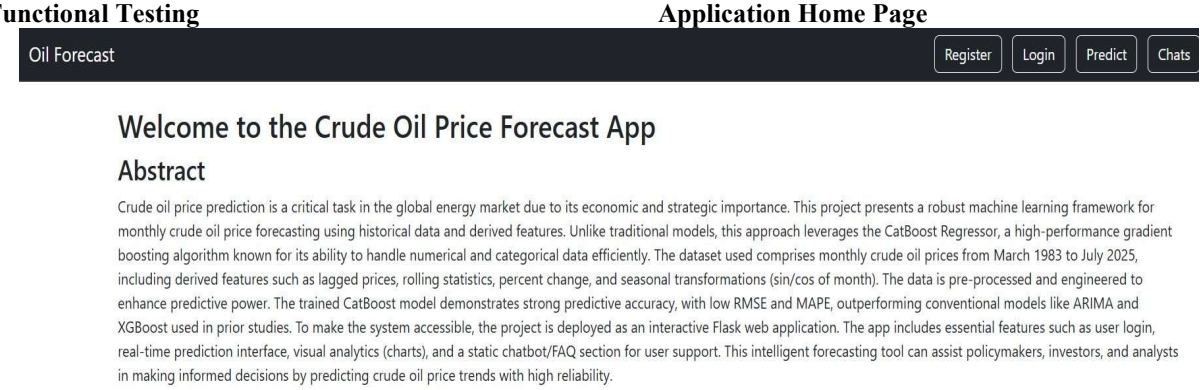
Acceptance Testing

End-to-end user workflow was validated — from registration and login through manual feature input to prediction result and chart display — confirming the system meets user expectations and functional requirements.

Snapshots

This chapter presents the actual screenshots from the implemented Crude Oil Price Forecasting web application, demonstrating the system's interface, functionality, and output visualizations.

Application Home Page



Home page of the Crude Oil Price Forecast Flask web application

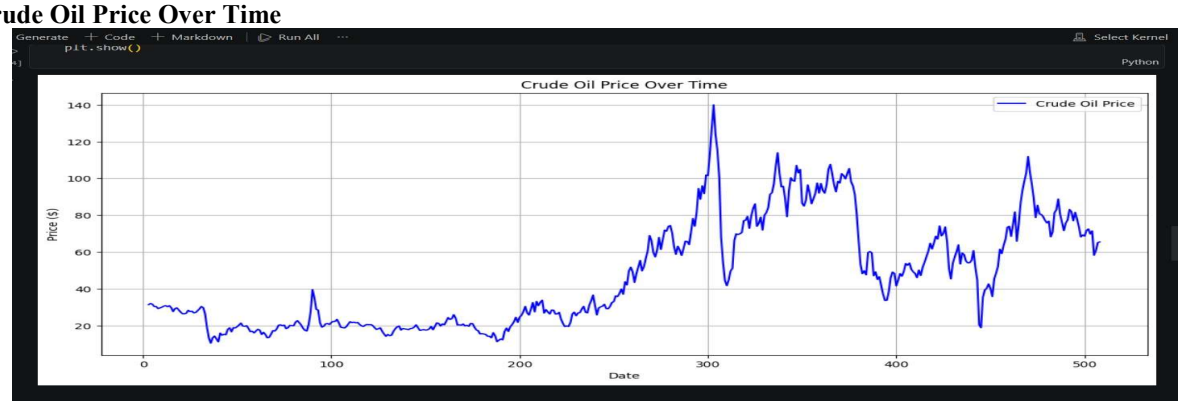
Source Code – app.py (Flask Backend)

https://doi.org/10.63665/IJMEC.1104s.42
ISSN: 2456-4265
IJMEC 2026

```python
from flask import Flask, render_template, request, redirect, url_for, session, flash
from catboost import CatBoostRegressor
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os

app = Flask(__name__)
app.secret_key = 'secret_key123'

model = CatBoostRegressor()
model.load_model("catboost_model.cbm")

users = {}

@app.route('/')
def home():
    return render_template("home.html")

@app.route('/register', methods=['GET', 'POST'])
def register():
    if request.method == 'POST':
        username = request.form['username']
        password = request.form['password']
        if username in users:
            flash('User already exists!')
        else:
            users[username] = password
            flash('Registration successful!')
            return redirect(url_for('login'))
    return render_template("register.html")

@app.route('/login', methods=['GET', 'POST'])
def login():
```

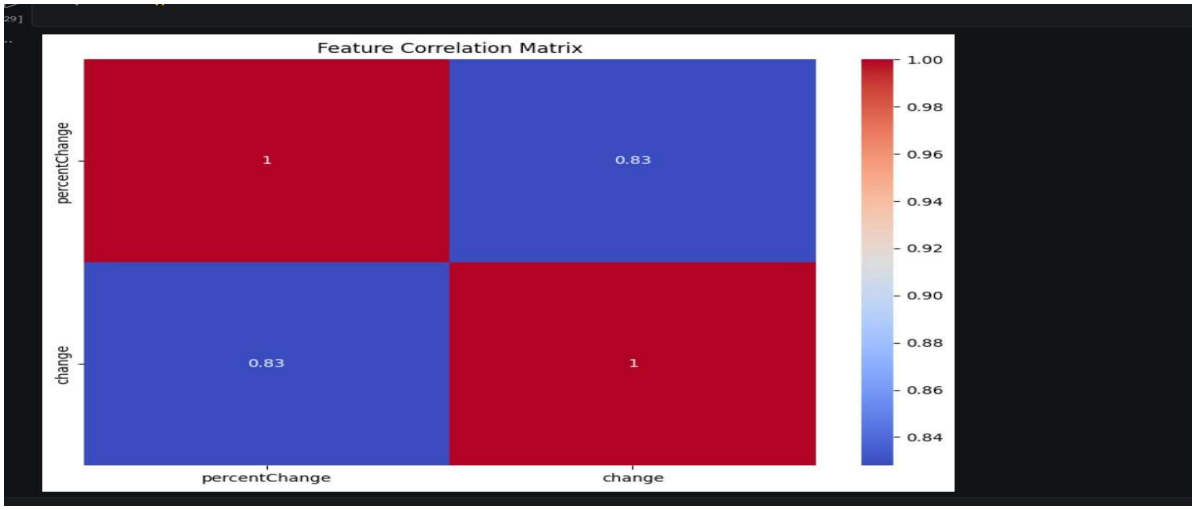
Flask backend source code showing CatBoost model loading, user registration and login routes

Crude Oil Price Over Time



Line chart showing crude oil price trend from 1983 to 2025

Feature Correlation Matrix



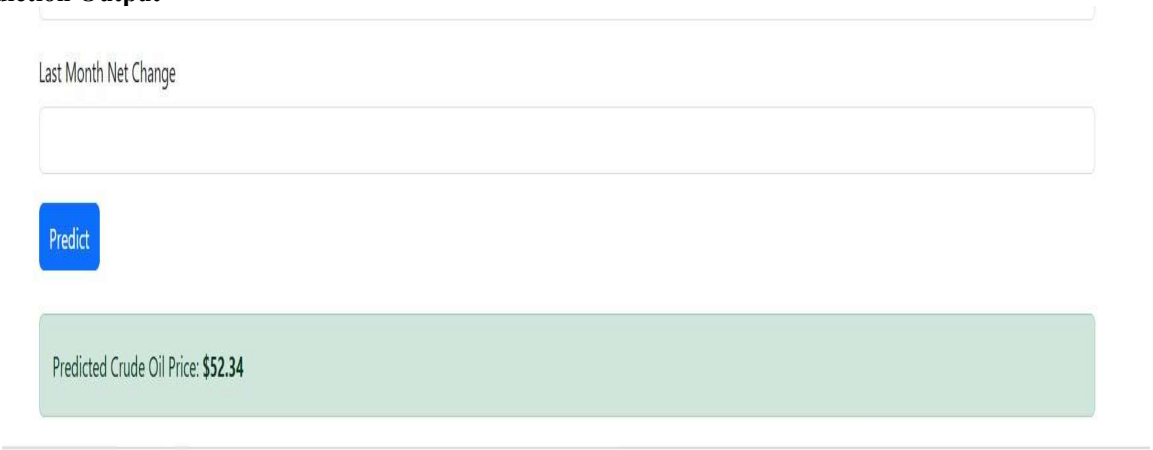
Heatmap showing correlation between percentChange and change features (correlation = 0.83)

Prediction Input Form

https://doi.org/10.63665/IJMEC.1104s.42
ISSN: 2456-4265
IJMEC 2026

Prediction form with input fields for lag prices, rolling statistics, month, year and percent change

Prediction Output



Model output showing predicted crude oil price of $52.34

Future Enhancements

The project can be extended in several meaningful directions:

Integration of external market indicators: global demand, production levels, OPEC policies, and stock market indices

Real-time data feeds to enable dynamic forecasting rather than relying only on static historical data

Deep learning exploration: LSTMs or Transformer-based models for capturing longer dependencies

Ensemble models combining CatBoost with other advanced algorithms for further performance gains

Interactive dashboards and automated reporting tools for policymakers and investors

Cloud-based deployment with REST APIs for seamless integration with financial and energy platforms

Conclusion

This project successfully demonstrates the application of machine learning techniques for forecasting monthly crude oil prices using historical data. By leveraging the CatBoost Regressor and carefully engineered time-series features, the system captures non-linear patterns, seasonal trends, and short-term fluctuations with high accuracy.

Unlike traditional statistical models, the proposed approach offers improved performance while reducing complexity, making it both scalable and practical. The evaluation metrics confirm the reliability of the model, and the inclusion of visualizations further enhances interpretability by providing insights into market behavior.

Overall, this project highlights the potential of machine learning in financial and energy forecasting, offering a data-driven solution that can support policymakers, investors, and industry stakeholders in

https://doi.org/10.63665/IJMEC.1104s.42
ISSN: 2456-4265
IJMEC 2026

making informed decisions about future oil price movements.

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, 2014.
[2] H. Mohammadi and L. Su, 'International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models,' Energy Econ., vol. 32, no. 5, pp. 1001–1008, 2010.
[3] T. Yao and Y.-J. Zhang, 'Forecasting crude oil prices with the Google index,' Energy Proc., vol. 105, pp. 3772–3776, 2017.
[4] H. Chiroma et al., 'Evolutionary neural network model for west Texas intermediate crude oil price prediction,' Appl. Energy, vol. 142, pp. 266–273, 2015.
[5] Y. Zhao, J. Li, and L. Yu, 'A deep learning ensemble approach for crude oil price forecasting,' Energy Econ., vol. 66, pp. 9–16, 2017.
[6] M. Wang et al., 'A novel hybrid method of forecasting crude oil prices using complex network science and artificial intelligence algorithms,' Appl. Energy, vol. 220, pp. 480–495, 2018.
[7] H. He et al., 'A novel crude oil price trend prediction method: Machine learning classification algorithm based on multi-modal data features,' Energy, vol. 244, 2022.
[8] R. Li et al., 'A novel multiscale forecasting model for crude oil price time series,' Technol. Forecasting Social Change, vol. 173, 2021.

https://doi.org/10.63665/IJMEC.1104s.42
ISSN: 2456-4265
IJMEC 2026