Progress Report – Predicting Future Oil and Natural Gas Prices

Due: April 11, 2025
Group Name: Data Pioneers

Members:
Widyan Hussien - coder
Ali Javaid - Project manager and code architecture design
Jamaima Syed - github manager
Alejandro Vega - researcher
Angela Marie Abrea - coder
Salma Abbady - absent

1. Introduction

Oil and natural gas prices are subject to frequent fluctuations driven by a wide range of factors, including supply-demand imbalances, geopolitical tensions, macroeconomic trends, and market speculation. This volatility presents significant challenges for policymakers, energy companies, and consumers alike.

In this project, we aim to develop a predictive system capable of forecasting short-term oil and gas prices, specifically focusing on Brent crude oil. By leveraging historical pricing data, macroeconomic indicators, and machine learning techniques, we seek to produce accurate daily or weekly price predictions that can inform better decision-making.

2. Literature Review - looking at what other experts have done

Recent efforts in oil and gas price prediction have explored a variety of methods, ranging from statistical models like ARIMA to more advanced machine learning techniques such as Random Forests and deep learning.

- **GitHub: Applied Data Science in Oil & Gas** – This repository introduces techniques such as time series forecasting and clustering tailored for oil industry data. The work focuses on exploratory analysis and modeling using Python-based tools, similar to our initial approach using regression and random forests.

- **SB Consulting Blog** – Discusses data science as a strategic tool for oil and gas decision-making. While not technical, it supports the relevance and business value of predictive systems in the energy sector.

- **Academic Sources (please confirm if you want me to add real citations here)** – Many academic papers have applied LSTM models for energy price prediction due to

their ability to capture sequential dependencies. Our approach plans to adopt similar models after establishing a solid baseline with simpler methods.

- **Forecasting crude oil price using LSTM neural networks** – This study developed a LSTM model to forecast crude oil prices. Additionally, they used an ANN model and an ARIMA model to compare and evaluate generalization ability, forecasting accuracy and stability, and timescale accuracy. The study uses data from WTI and Brent, starting from February 1986 to May 2021. Similarly to our approach, this study employed a LSTM model and used Brent crude oil prices. Additionally, this study highlights the effectiveness of LSTM models compared to others, which supports our choice of using LSTM. However, it differs by not incorporating additional factors, such as macroeconomic indicators, which we hope will increase our model's accuracy.
- **Crude oil price forecasting using K-means clustering and LSTM model enhanced by dense-sparse-dense strategy** – This study also uses a LSTM neural network, however it combines the model with K-means clustering. Additionally, a DSD technique is also used in this study's model, creating a model with a three-step training process. Like the previous study, researchers also used both WTI and Brent prices. As both studies we observed used WTI in addition to Brent, perhaps it wouldn't hurt to include the WTI dataset in our approach. This study highlighted the improved performance of a model that combines various techniques. In addition to increased accuracy, the study notes the model also yielded a faster deployment speed.

3. Dataset

We are currently using the following datasets:

- **Brent Crude Oil Prices Dataset** (from Kaggle):
    - Source: Kaggle Dataset Link
    - Timeframe: 1987–2021
    - Size: ~9,000 daily entries
    - Key fields: Date, Price
- **Macroeconomic Indicators**:
    - Inflation and interest rates (Federal Reserve Economic Data – FRED)
    - S&P 500 Index (Yahoo Finance)
    - US Crude Oil Inventories and Production (EIA)

- **Crude Oil Price** (WTI, from Kaggle):
    - Source: https://www.kaggle.com/datasets/sc231997/crude-oil-price/data
    - Timeframe: March 1983 - March 2025
    - Size: ~500 monthly entries
    - Fields: date, price, %change, change

Preprocessing Steps, below are our plans that we will use to do the calculations:

Step	Explanation
Handling Missing Values	Forward fill (use previous value), drop rows, or use mean imputation
Normalization / Scaling	Apply Min-Max or Z-score normalization on features
Time Windowing	Convert data into sliding windows (e.g., last 30 days as input, next day as output)
Feature Engineering	Create lag features, moving averages, etc.
Merging Data	Join oil prices with S&P 500 or inflation data by date

4. Baseline Model

To evaluate the performance of our predictive system, we implemented a naive baseline model which assumes that the oil price for the next day is the same as the current day's price:

y^t+1=yt

This model serves as a simple benchmark and represents a lower bound for performance. Despite its simplicity, it often performs surprisingly well in financial time series due to high short-term autocorrelation.

We evaluated the model on a 30% held-out test set using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The results are as follows:

Model	RMSE	MAE
Naive Baseline	2.41	1.78

5. Main Approach

We started with two models:

1. Linear Regression

- *Inputs: Past 30 days of prices + inflation + interest rate + S&P 500 index*
- *Output: Predicted next-day price*

2. Random Forest Regressor

- *Handles nonlinear interactions and performs well on tabular data*
- *Tuned for 100 estimators*

Results:

Model	*RMSE*	*MAE*
Linear Regression	*1.96*	*1.52*
Random Forest	*1.44*	*1.13*

These results show an improvement over the naive model, with Random Forest currently outperforming Linear Regression.

6. Evaluation Metrics

We use:

- **Root Mean Squared Error (RMSE)**:

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

- **Mean Absolute Error (MAE)**:

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

These metrics are effective in capturing both average and extreme prediction errors.

7. Results & Analysis

At this stage of the project, we have designed and planned our baseline and initial machine learning models, but have not finalized or fully implemented them yet. However, we have outlined the expected outcomes, evaluation plan, and areas we intend to analyze once results are available.

*We anticipate that the **naive baseline model**, which simply assumes that the next day's oil price will be equal to the current day's price, will serve as a useful lower-bound benchmark. Based on literature and preliminary data patterns, we expect this model to perform reasonably well in very short-term forecasts due to the autocorrelated nature of time series financial data.*

*For our **main approaches**, including Linear Regression and Random Forest Regressor, we hypothesize that incorporating multiple economic indicators (such as inflation, S&P 500, and oil inventory levels) will improve performance over the naive baseline. These models are expected to better capture nonlinear dependencies and interactions between variables, which may be critical in predicting price fluctuations influenced by macroeconomic shifts.*

Once the models are implemented and evaluated using RMSE and MAE on a held-out test set, we plan to conduct the following types of analysis:

- ***Performance comparison** between baseline and machine learning models.*
- ***Feature contribution analysis** to understand which economic variables most influence predictions.*

- ***Temporal error trends*** *to assess model stability across different time periods (e.g., volatile vs. stable market windows).*

- ***Visualization*** *of actual vs. predicted price trends over selected time intervals.*

We will also explore potential cases where models underperform — such as during geopolitical shocks or sudden market changes — and use these insights to guide refinement of the model architecture or features.

Further analysis will follow after completing initial experiments and gathering quantitative results in the next stage of development

8. References

Jahandoost, A., Abedinzadeh Torghabeh, F., Hosseini, S. A., & Houshmand, M. (2024). Crude oil price forecasting using K-means clustering and LSTM model enhanced by dense-sparse-dense strategy. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00977-8

Zhang, K., & Hong, M. (2022). Forecasting crude oil price using LSTM Neural Networks. *Data Science in Finance and Economics*, *2*(3), 163–180. https://doi.org/10.3934/dsfe.2022008