Predicting Oil Spot price using Time Series Data

Richie Zhang rlz6628@nyu.edu
Uday Sapra us2055@nyu.edu
Rohan Sanghrajka rjs9346@nyu.edu

Milestone 2

https://github.com/RichieZrq/Deep-Learning-Oil-Price-Prediction-Project

1. Research Question

What are the factors that influence oil spot prices, and can we predict these prices using time series analysis combined with external asset classes and macroeconomic indicators?

2. Methodology

2.1 Feature Selection

Each selected feature represents a critical dimension influencing the oil market, whether through direct pricing, macroeconomic sentiment, or risk assessment. By incorporating these variables, we aim to capture a multi-dimensional view of oil price determinants. Here's a breakdown of the rationale behind each feature:

1. **S&P 500 Index:** As a barometer of economic health, the S&P 500 reflects investor sentiment and economic activity levels, impacting oil demand and pricing. A robust or weak S&P 500 can indicate trends in economic expansion or contraction, respectively, which directly correlates with energy demand.

2. **Interest Rate:** Interest rates set by central banks influence the cost of borrowing and capital investment, which impacts economic growth. Higher interest rates typically discourage borrowing, slowing down economic activity and oil consumption. Conversely, lower interest rates stimulate economic growth, leading to increased energy needs.

3. **VIX (Volatility Index):** Known as the "fear gauge," VIX measures market volatility. High volatility may deter investments in risk-sensitive sectors, including commodities like oil, thereby affecting demand. Tracking VIX alongside oil prices helps anticipate fluctuations in investor sentiment and risk appetite.

4. **OVX (Oil Volatility Index):** Specific to crude oil, OVX captures market expectations of future volatility in oil prices, making it a direct indicator of risk in the oil market. Elevated OVX values suggest anticipated instability in oil prices, which is useful for understanding potential price swings.

5. **USO (United States Oil Fund):** USO is a commodity ETF that closely tracks WTI crude oil prices. By monitoring its performance, we gain insight into market

participants' collective sentiment and flows within the oil sector, providing a proxy for real-time oil demand and supply factors.

6. **DXY (US Dollar Index):** Oil prices are inversely related to the strength of the USD, as oil is typically priced in dollars. A stronger dollar makes oil more expensive for non-USD countries, potentially reducing global demand, while a weaker dollar can boost demand.

7. **Crude Oil Historical Data:** Including crude oil's historical spot prices enables us to observe patterns, seasonality, and autocorrelations essential for accurate modeling. Historical data acts as a baseline for evaluating the effects of new features, especially in detecting time-dependent trends.

Our hypothesis suggests that interest rates will have the highest impact on oil prices, as they directly influence economic growth and investment, driving changes in oil demand. Historical crude oil data is also expected to have a significant impact, providing essential insights into intrinsic patterns and trends crucial for short-term price forecasting. USO, which tracks WTI crude prices, is anticipated to show a strong positive correlation with oil demand, reflecting immediate market sentiment. The DXY (US Dollar Index) is predicted to have a moderate inverse effect, as a stronger dollar makes oil pricier globally, potentially curbing demand. The OVX, specific to oil price volatility, is expected to impact short-term price fluctuations, capturing risk sentiment directly related to oil. The S&P 500 should have a moderate effect, reflecting overall economic health and influencing oil consumption trends. Finally, the VIX is expected to have a smaller, indirect influence, with higher volatility often signaling reduced risk appetite across assets, including oil. Collectively, these features should provide a well-rounded perspective for forecasting oil price movements.

2.2 Scraping

We scraped OPEC press releases using Selenium. Although the original plan was to use request and beautiful soup, the requests were consistently blocked by websites, forcing us to pivot to using Selenium to simulate a real user. As a result, each press release page between 2021 January and 2024 October is stored as a single string in a column named text, and the date is stored separately in a column named date.

2.3 Sentiment Analysis using BERT models

Using the scraped dataset, we applied multiple pretrained BERT models including vanilla BERT, ALBERT, FinBERT fine-tuned on financial articles, and CrudeBERT fine-tuned on oil related headlines. For each model, we loaded the pretrained model, and defined the tokenizer using transformers.AutoTokenizer. Each sentence is then tokenized using the models' specific tokenizer, and the model performs inference to obtain logits for three sentiment classes: negative, neutral, and positive. These logits are passed through a softmax layer to produce class probabilities, which are then used to calculate a sentiment score on a scale from -1 to 1, where negative scores indicate negative sentiment, and positive scores indicate positive sentiment. Finally, we computed the average sentiment scores of all sentences to provide an overall sentiment score for the text. After experimentation, we found that the optimal model is CrudeBERT since the other models failed to output results that consistently matched the sentiment of the different texts. That being said, since the texts from OPEC are all professional and maintained a neutral tone, predictions from BERT models indicated that the press releases are generally neutral or negative

2.4 Sentiment Analysis using LLMs

Since the BERT output was mostly neutral, we felt it was necessary to understand the news and its sentiment in context to oil markets and prices. The announcement may have a neutral tone in its general language, however, it could still have a positive and negative sentiment toward oil prices and markets. Hence, for our purpose, a contextual understanding of oil news is necessary. Such contextual understanding of news could be introduced by using pre-trained Large Language models. Further, with Prompt Engineering, we can introduce task instructions and relevant context. We tested two open-source LLMs - Llama 3.1 8b and Mistral NeMo 12B. Llama was chosen because it is an industry-standard for open-source models. Mistral's NeMo is a newer model that has better benchmarking scores than Llama 3.1 8b, it is also slightly larger. These were chosen because they show good accuracy while balancing size and compute needs. Given limited GPU compute and the straightforward nature of our task, these seemed sufficient and we did not test larger Llama models - 70B or 405B.

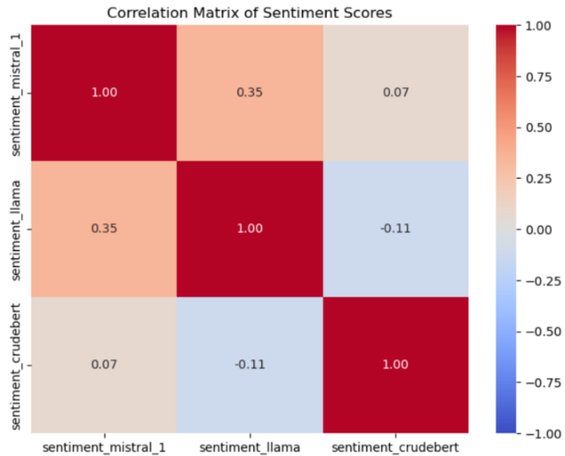
2.5 Modeling

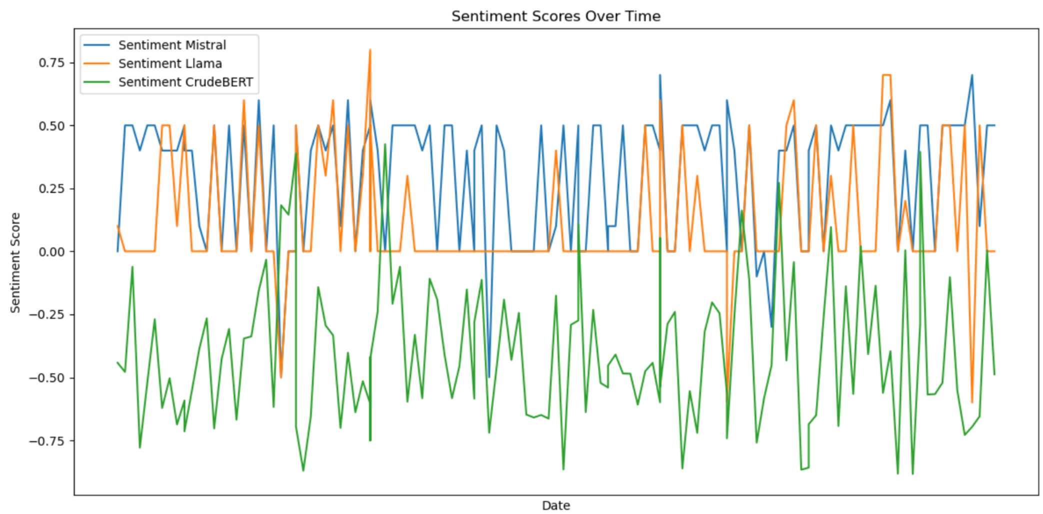
We prototyped our baseline transformers model - PatchTST. It is a transformers-based time-series forecasting model that excels at univariate forecasting tasks. For this autoregressive model, we only needed one feature - spot oil prices. Since we are still working on feature selection and feature engineering, the multivariate model (Temporal Fusion Transformer - TFT) will be developed later. PatchTST will be our baseline transformer forecasting model and we will study the information and accuracy gain we have from incorporating exogenous variables with the TFT. For PatchTST development,we first trained the model with the hyperparameters used in the original paper. To further improve results for our data, we tuned the hyperparameters with 400 runs on Optuna to optimize for validation-set mse.

3. Results

3.1 Sentiment Analysis

After obtaining sentiment scores using the three methods (Llama, Mistral, and CrudeBERT), we computed the correlation between OPEC sentiment scores of the three methods and created a visualization of all sentiment scores to understand them:



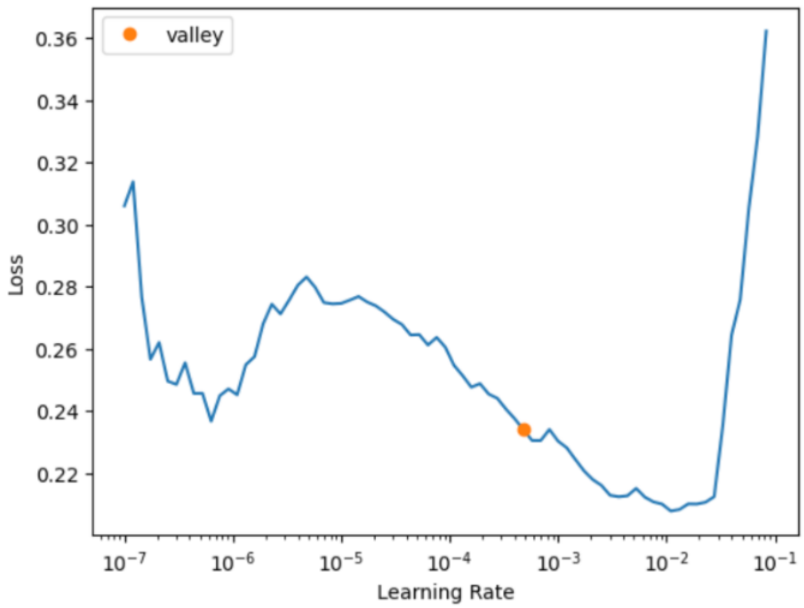


3.2 Modeling

For PatchTST, below are the ideal set of hyper-parameters we obtained from our Optuna runs using the TPESampler.

Params:

`fcst_history:` 50 - the context the model gets to make each set of future predictions
`fcst_horizon:` 16 - the number of future predictions the model makes given its set of context
`batch_size:` 20
`epochs:` 14
`n_layers:` 4 - number of hidden layers
`n_heads:` 2 - number of attention heads per layer
`d_model:` 112 - dimension of attention layers
`d_ff:` 224 - dimension of fully connected layers
`attn_dropout:` 0.2
`dropout:` 0.30000000000000004
`patch_len:` 11 - number of tokens/time-steps per input patch
`stride:` 2



The learning rate bounds of the model were set by training models at different lr's and then identifying the values of `lr_min` and `lr_max`. Then a cyclical `cycle_policy` was used for model training.

The final MSE on val-set we obtained was 0.0739. The test-set MSE was also in the range of 0.06-0.07.

4. Analysis and Plans for Additional Analysis

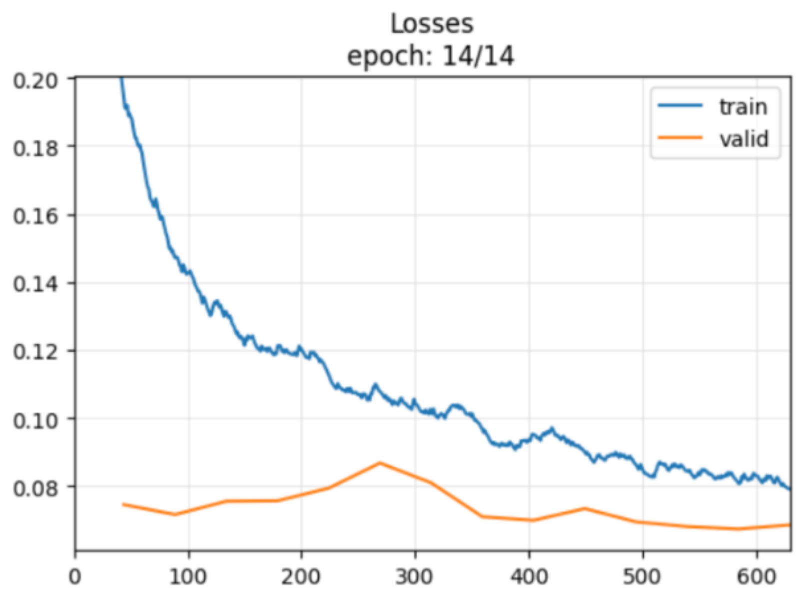
4.1 Sentiment Analysis

All three models provided different sentiment scores, as revealed by the low correlations and diverse outputs. In general, CrudeBERT outputs negative sentiment scores for OPEC press releases, which is possibly due to the professional and sober tone across all articles. On the other hand, the two LLMs output more similar scores, shown from their relatively higher correlation. Both output rather positive scores in general, with Llama being relatively more neutral. The variability in results were somewhat expected, given that each model is trained on different data or uses different fine-tuning methods for sentiment detection.

That being said, we determined that the OPEC press releases were not the best source for sentiment analysis as they were all kept neutral tones regardless of the events. We will move on to scraping and analyzing sentiment scores for multiple news headlines from 2021 January to now. We plan to use different models for sentiment analysis like how we experimented using OPEC articles. At the end, we will select the optimal model and use sentiment scores as a feature in our final multivariate model.

4.2 Modeling

The model training was very stable across epochs. The loss training loss and validation loss decreased throughout epochs - yet, at all times, the validation set loss was below the training-set loss. This highlights the impressive generalizability of the model and stable learning under the selected hyperparameter-set. Our true analysis will begin once we run our multivariate forecasting model with out entire feature-set and compare that with the autoregressive PatchTST. We also plan to add ARIMA/SARIMA/ARIMA for purely benchmarking purposes in the end.



5. Work Plan

October (completed)

1. **Data Collection and Preprocessing**: Rohan

2. **Feature Engineering**

   (a) Hypothesis (Finance & Commodities Domain Knowledge) - Rohan

   (b) Data Scraping - Richie

   (c) Sentiment Analysis - Richie and Uday

   (d) Implementation - Uday

3. **Hypothesis & Correlation Testing and Feature Selection**: Uday

November

4. **Model Hypothesis & Correlation Testing and Feature Selection** - Rohan

5. **Model Prototyping, Development, and Validation (TFT and PatchTST)** - Uday
6. **Benchmarking Model Prototyping, Development, and Testing (ARIMA and Sarima)** - Richie and Rohan
7. **News Scraping and Sentiment Analysis** - Richie

December

6. **Results and Conclusions (Visualizations)** - Rohan and Uday
7. **Presentation and Report** - Rohan, Uday, and Richie
8. **Deployment and Twitter API Integration for Automatic Prediction Updates** - Richie