Quantitative Analysis and Forecasting of Crude Oil Prices with Integrated Risk Management

Ishan Banerjee

May 4, 2025

Abstract

This report documents a comprehensive project to forecast next-day returns for WTI crude oil using a machine learning model and to implement a simple trading strategy. The process includes data collection from Yahoo Finance, preprocessing (flattening multi-index columns, computing returns, and feature engineering), training a Random Forest model, evaluating its performance, and backtesting a trading strategy. Final outputs include model performance metrics, feature importance analysis, and trading strategy performance metrics.

1 Introduction

Crude oil is a key commodity that is influenced by various factors such as geopolitical events, macroeconomic indicators, and market sentiment. The objective of this project is to develop a data-driven approach to forecast next-day returns for WTI crude oil and to implement a trading strategy based on these predictions. The methodology combines data preprocessing, machine learning using a Random Forest model, and the backtesting of a simple long/short strategy.

2 Data Collection and Preprocessing

2.1 Data Sources

Data was collected from Yahoo Finance for the following instruments:

- **WTI Crude Oil Futures (CL=F)**
- **US Dollar Index (DX-Y.NYB)**
- **10-Year Treasury Yield (T̂NX)**

The data spans from January 1, 2018, to the present.

2.2 Flattening and Renaming Columns

Yahoo Finance sometimes returns data with a MultiIndex. We flattened the columns and renamed them to intuitive names. The final results are:

- **Flattened Columns:** `['WTI_Close', 'USD_Index_Close', 'TenYrYield_Close', 'WTI_Returns', 'USD_Returns', 'TenYrYield_Change', 'WTI_Future_Return', 'WTI_Returns_ 'USD_Returns_lag1', 'TenYrYield_Change_lag1', 'WTI_Returns_lag2', 'USD_Returns_lag2', 'TenYrYield_Change_lag2', 'WTI_Returns_lag3', 'USD_Returns_lag3', 'TenYrYield_Change_`
- **Renamed Columns:** `['WTI_Close', 'USD_Index_Close', 'TenYrYield_Close', 'WTI_Returns', 'USD_Returns', 'TenYrYield_Change', 'WTI_Future_Return', 'WTI_Returns_ 'USD_Returns_lag1', 'TenYrYield_Change_lag1', 'WTI_Returns_lag2', 'USD_Returns_lag2', 'TenYrYield_Change_lag2', 'WTI_Returns_lag3', 'USD_Returns_lag3', 'TenYrYield_Change_`

2.3 Computing Returns and Feature Engineering

Daily returns for each series were computed as follows:

$$\begin{aligned}
\text{WTI_Returns} &= \frac{\text{WTI_Close}_t - \text{WTI_Close}_{t-1}}{\text{WTI_Close}_{t-1}}, \\
\text{USD_Returns} &= \frac{\text{USD_Index_Close}_t - \text{USD_Index_Close}_{t-1}}{\text{USD_Index_Close}_{t-1}}, \\
\text{TenYrYield_Change} &= \frac{\text{TenYrYield_Close}_t - \text{TenYrYield_Close}_{t-1}}{\text{TenYrYield_Close}_{t-1}}.
\end{aligned}$$

The target variable is defined as the next-day return of WTI:

$$\text{WTI\_Future\_Return} = \text{WTI\_Returns}_{t+1}.$$

In addition, lagged features (1, 2, and 3 days) were created for each return series to capture temporal dependencies.

2.4 Visualizations

Figure 1 displays the time series of the three price series.

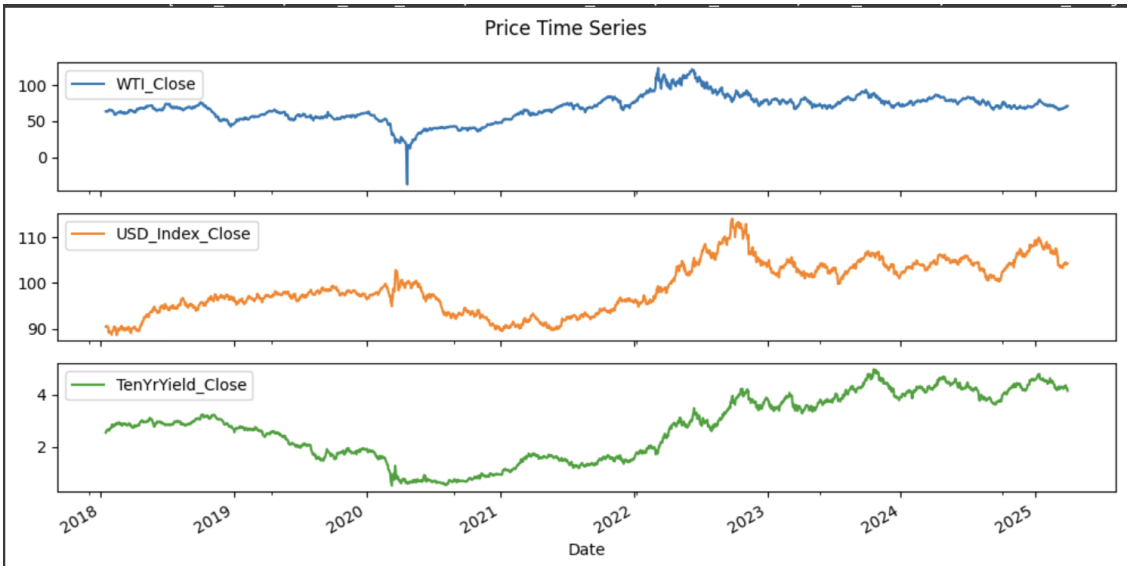

Figure 1: Time series of WTI crude oil, US Dollar Index, and 10-Year Treasury Yield.

Figure 2 shows the correlation heatmap among daily returns and lagged features.

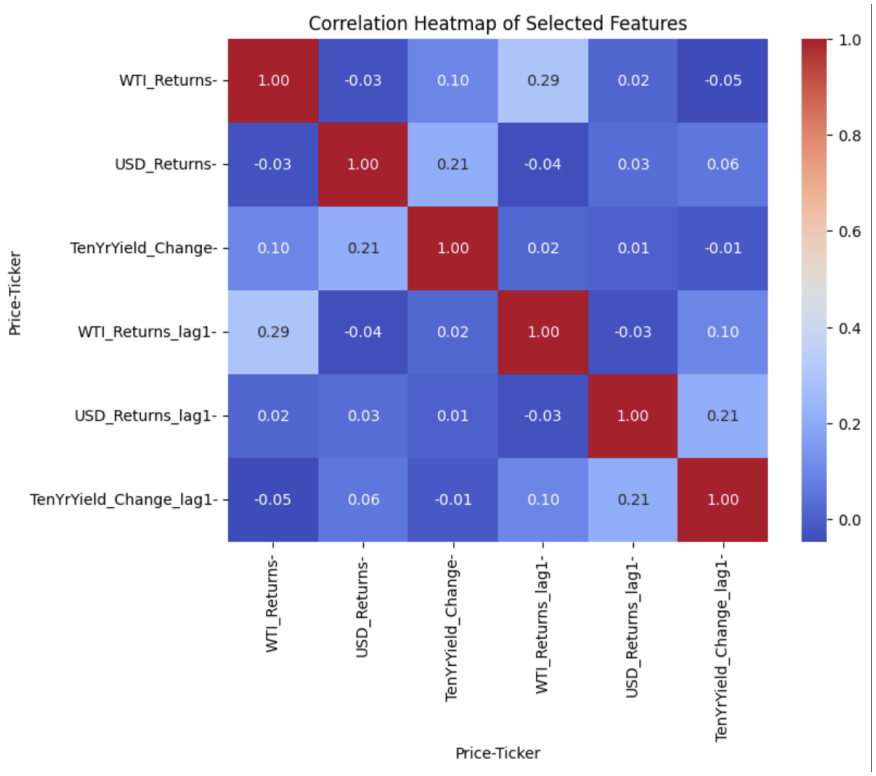

Figure 2: Correlation heatmap of selected features.

2.5 Final Preprocessed Data Sample

The final DataFrame includes the following columns:

- `WTI_Close, USD_Index_Close, TenYrYield_Close`
- `WTI_Returns, USD_Returns, TenYrYield_Change`
- `WTI_Future_Return` (target variable)
- Lagged features: `WTI_Returns_lag1, USD_Returns_lag1, TenYrYield_Change_lag1, WTI_Returns_lag2, USD_Returns_lag2, TenYrYield_Change_lag2, WTI_Returns_lag3, USD_Returns_lag3, TenYrYield_Change_lag3`

A sample of the final data is shown in Table 1.

Date	**WTI_Close**	**USD_Index_Close**	**TenYrYield_Close**	**WTI_Returns**	**USD_Retur**
2018-01-16	63.73	90.39	2.544	-0.00887	-0.00638
2018-01-17	63.97	90.54	2.578	0.00377	0.00166
2018-01-18	63.95	90.50	2.611	-0.00031	-0.00044
2018-01-19	63.37	90.57	2.637	-0.00907	0.00077
2018-01-22	63.49	90.40	2.665	0.00189	-0.00188

Table 1: Sample of the final preprocessed data. (Additional columns include lagged features.)

3 Machine Learning Model: Random Forest Approach

3.1 Target and Feature Setup

The target variable is defined as `WTI_Future_Return`, the next-day return of WTI crude oil (obtained by shifting `WTI_Returns` by one day). The features include the current-day returns and 1, 2, and 3-day lagged values for WTI, USD, and 10-Year Treasury Yield.

3.2 Train/Test Split

To avoid look-ahead bias, a time-based split was used:

- **Training Set:** Data before 2022-01-01.
- **Test Set:** Data from 2022-01-01 onward.

3.3 Model Training and Evaluation

A Random Forest Regressor with 200 trees was trained on the training set. The evaluation on the test set produced the following results:

- **Test MSE:** 0.000883
- **Test R^2:** -0.6233

Figure 3 shows a plot comparing actual and predicted next-day returns.

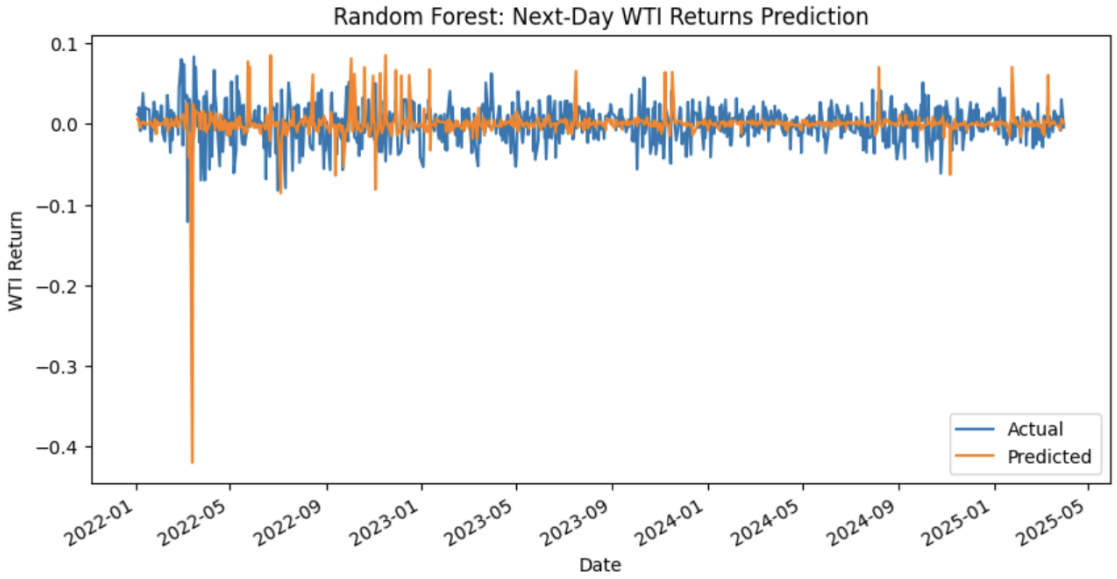

Figure 3: Actual vs. Predicted next-day WTI Returns using Random Forest.

3.4 Feature Importance Analysis

Table 2 shows the feature importances obtained from the Random Forest model.

Feature	Importance
WTI_Returns	0.2926
TenYrYield_Change_lag2	0.1663
WTI_Returns_lag3	0.1048
USD_Returns_lag2	0.0772
WTI_Returns_lag1	0.0583
TenYrYield_Change	0.0691
WTI_Returns_lag2	0.0516
USD_Returns_lag1	0.0480
TenYrYield_Change_lag3	0.0428
USD_Returns_lag3	0.0339
USD_Returns	0.0297
TenYrYield_Change_lag1	0.0283

Table 2: Feature Importances from the Random Forest model.

4 Trading Strategy and Backtesting

A simple trading strategy was implemented based on the model's predictions:

- If the predicted next-day return is positive, a long position (+1) is taken.
- If the predicted next-day return is negative, a short position (−1) is taken.

The trading signal is shifted by one day to simulate trading on the following day. Daily strategy returns are calculated as the product of the actual next-day return and the trading signal.

Figure 4 shows the cumulative return of the strategy over the test period.

Figure 4: Cumulative Return of the Trading Strategy based on Random Forest predictions.

The performance metrics of the strategy are:

- **Final Strategy Return:** 99.53% over the test period.
- **Strategy Sharpe Ratio:** 0.76.

5 Conclusion

This project demonstrated a complete pipeline for forecasting next-day returns for WTI crude oil and implementing a trading strategy based on the predictions:

1. **Data Preprocessing:** Data was collected from Yahoo Finance, the MultiIndex columns were flattened and renamed, and daily returns along with lagged features were computed.
2. **Machine Learning Modeling:** A Random Forest model was trained to predict next-day returns. Although the test R^2 was negative, the model provided actionable signals.
3. **Trading Strategy:** A simple long/short trading strategy based on the model's predictions achieved a cumulative return of 99.53% and a Sharpe Ratio of 0.76 over the test period.

6 Future Work

Potential future enhancements include:

- Exploring alternative machine learning models such as XGBoost or LSTM networks.
- Incorporating additional macroeconomic factors (e.g., inflation rates, inventory data).
- Refining the trading strategy with advanced risk management techniques (e.g., Value-at-Risk, stress testing).

References

- Yahoo Finance, `https://finance.yahoo.com/`
- Scikit-learn Documentation, `https://scikit-learn.org/`
- Pandas Documentation, `https://pandas.pydata.org/`