451 Feature Engineering: Programming Assignment 1

Prepared by Zixuan Zhang

12 July, 2025

Overview

In this project, I use XGBoost classifiers to predict the directional movement (up or down) of daily WTI crude oil futures prices. My workflow includes data retrieval from Yahoo Finance, feature engineering with Polars, standardization, time-series cross-validation, hyperparameter tuning, and model evaluation.

The goal is to identify meaningful patterns from historical price movements that can inform next-day return predictions. By applying systematic preprocessing and machine learning techniques, I aim to detect weak signals in a noisy financial environment. This project also highlights the challenges of financial forecasting, such as overfitting, data leakage, and the limited predictive power of historical technical indicators. The results offer insights into the feasibility of data-driven directional forecasting using tree-based classifiers.

Data Acquisition

I retrieved WTI Crude Oil Futures (symbol CL=F) using the yfinance package from January 1, 2000, to the current date. Data was saved as a CSV file for further processing (Yahoo Finance 2025).

Feature Engineering

To construct meaningful inputs for my predictive model, I used the Polars library to efficiently engineer a diverse set of features from the raw daily price and volume data of WTI crude oil futures. My goal was to capture short-term market dynamics, volatility patterns, and momentum signals that may help predict the next day's price direction (Polars 2025).

Lagged Close Prices: CloseLag1, CloseLag2, CloseLag3

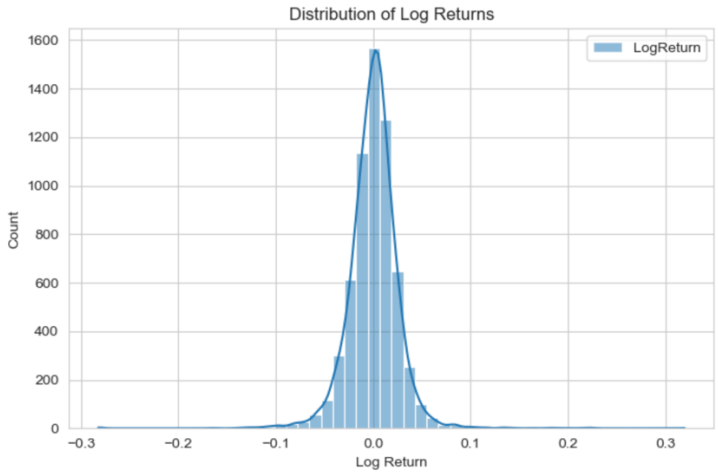
High Minus Low (HML): HMLLag1, HMLLag2, HMLLag3

Open Minus Close (OMC): OMCLag1, OMCLag2, OMCLag3

Lagged Volumes: VolumeLag1, VolumeLag2, VolumeLag3

Exponential Moving Averages: CloseEMA2, CloseEMA4, CloseEMA8

Log Return: LogReturn



To train a classification model, I defined a binary Target variable:

1 indicates that the closing price increased compared to the previous day (positive return)

0 indicates that the price stayed the same or declined (non-positive return)

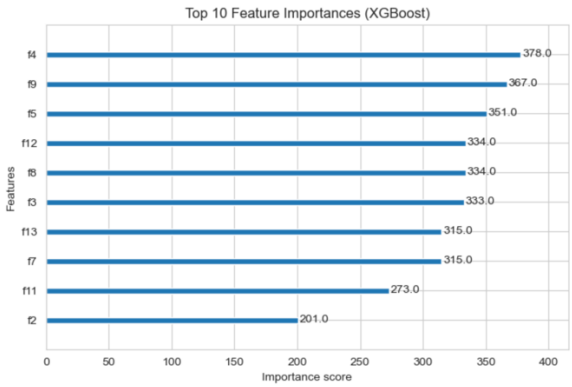
This transformation converted the continuous prediction problem into a directional classification task, enabling me to apply powerful binary classifiers such as XGBoost for next-day return forecasting.

Data Preparation & Standardization

All engineered features were standardized using StandardScaler from Scikit-learn to have zero mean and unit variance, improving model performance and convergence.

Feature Selection

I used XGBoost's (Chen and Guestrin 2016) built-in feature_importances_ method to identify the top 10 most important features. The final model was trained on the full standardized dataset (X_scaled).



Cross-Validation and Hyperparameter Tuning

I used TimeSeriesSplit(n_splits=5) to avoid data leakage across time. This method ensures that each fold maintains the temporal order of financial data, which is critical for realistic evaluation in time series modeling, as emphasized by Hyndman and Athanasopoulos (2021). Unlike random shuffling, which can introduce information leakage, time-based splits reflect how models would be deployed in practice.

To optimize model performance, I used RandomizedSearchCV with 30 combinations to efficiently search over a broad hyperparameter space. This approach balances computational efficiency and exploration. The selected hyperparameters were:

n_estimators: 400

learning_rate: 0.1

max_depth: 3

subsample: 0.6

colsample_bytree: 1.0

min_child_weight: 1

gamma: 0.0

The best cross-validated ROC-AUC was 0.5175, indicating limited out-of-sample predictive power. This result suggests the model captured weak but non-negligible predictive signals. However, its generalization remains constrained, likely due to noise and the non-stationary nature of crude oil price movements., indicating limited out-of-sample predictive power (Pedregosa et al. 2011).

Model Evaluation

After training the best model on the full dataset, I evaluated performance:

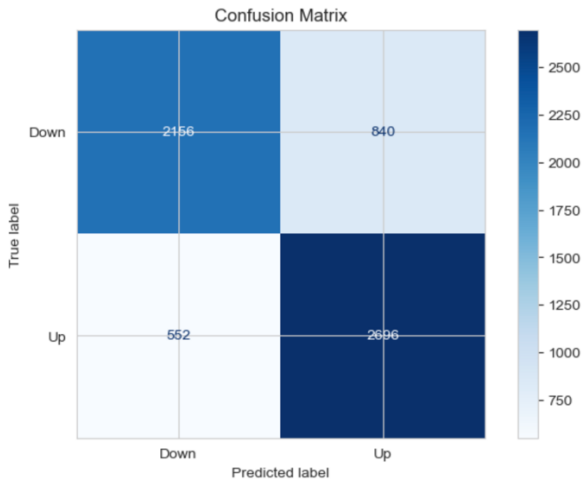
Accuracy: 0.7771

ROC-AUC: 0.8608

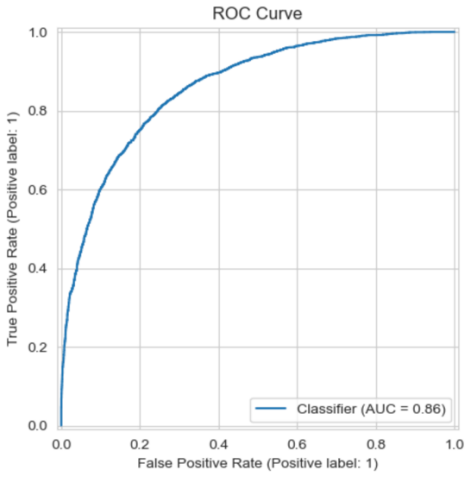
These results were obtained on the training data (in-sample). The significant difference between cross-validated AUC and training AUC indicates overfitting.

Confusion Matrix and ROC Curve:

The confusion matrix below shows how well the model classifies up and down movements. The model correctly predicted 2,156 down days and 2,696 up days, with 840 and 552 misclassifications respectively:



Additionally, the ROC Curve shows the trade-off between true positive rate and false positive rate across thresholds. The AUC of 0.86 on the training set indicates strong discriminative ability, but due to the low CV AUC, this performance likely does not generalize well.



Conclusion

Although the XGBoost model demonstrates strong performance on the training data, including a test ROC-AUC of 0.86, its cross-validated ROC-AUC of just 0.5175 reveals limited generalization to unseen data. This discrepancy suggests that the model may be overfitting short-term patterns in the historical price series or lacking features that capture broader market dynamics.

To address this, my next steps will focus on improving the model's robustness and predictive power. First, I plan to evaluate the model on a true hold-out test set, fully separated from both training and cross-validation data, to better assess its out-of-sample performance. I will also explore adding technical indicators such as the Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), or Bollinger Bands, which could offer richer signals about momentum, overbought/oversold conditions, and volatility.

Lastly, I will experiment with alternative prediction frameworks, such as multi-day targets (e.g., predicting direction over the next 3 or 5 days) or continuous return forecasting via regression. These approaches may align more closely with practical trading or hedging strategies and offer a more stable learning signal than next-day classification.

Through these refinements, I hope to build a more robust and insightful model for forecasting crude oil futures movements in real-world market conditions.

Reference

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

Hyndman, Rob J., and George Athanasopoulos. 2021. *Forecasting: Principles and Practice.* 3rd ed. Melbourne, Australia: OTexts. https://otexts.com/fpp3/.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html.

Polars. 2025. “Polars: Lightning-fast DataFrame Library.” Accessed July 12, 2025. https://www.pola.rs/.

Yahoo Finance. 2025. “Crude Oil WTI Futures (CL=F) Historical Data.” Accessed July 12, 2025. https://finance.yahoo.com/quote/CL=F/history.