Multi-Time Series Averaging of Ensemble Machine Learning Models Towards Crude Ol Price Forecasting

Farid Javadnejad, Ph.D.[1]

1. Introduction

Crude oil and other refined liquid products from fossil fuels are critical contributors to the world economy. Petroleum has been the largest energy source for all countries. Its products run vehicles, heat buildings, and produce electricity. Moreover, various industries use petroleum as a raw material to produce intermediate or end-user products that we use daily (EIA 2022b; Lu et al. 2021; Deng, Ma, and Zeng 2021; Kilian and Murphy 2014). In 2019, global petroleum consumption neared 100 million barrels per day (Table 1).

Table 1. The largest oil consumers and their share of total world consumption (EIA 2022b)

Ranking	Country	Million barrels per day	Share of world total
1	United States	20.54	20%
2	China	14.01	14%
3	India	4.92	5%
4	Japan	3.74	4%
5	Russia	3.70	4%
6	Saudi Arabia	3.18	3%
7	Brazil	3.14	3%
8	Canada	2.63	3%
9	South Korea	2.60	3%
10	Germany	2.35	2%
	World total	100.23	

[1] Geospatial Consultant, Aerogeospatial LLC, 1401 21st Street Suite R, Sacramento, CA 95811, e-mail: nejad.fj@gmail.com.

- *Problem statement*

Crude oil prices are difficult to predict accurately due to the number of influencing factors and the highly complex behavior of such influences. Global economic and social activities can be substantially impacted by fluctuations in crude oil prices. Threfore, despite challenges for prediction of oil price, accurate oil price forecasting is crucial for decision-making support for the manufacturing, logistics, and government sectors to guide industrial and social policies and practices (Kilian and Murphy 2014; Deng, Ma, and Zeng 2021; Lu et al. 2021).

- *Background*

Brent, West Texas Intermediate (WTI), Dubai/Oman, and Shanghai crude oil prices are the major benchmarks of the crude oil market and are reported is USD per barrel unit. Factors such as supply and demand, financial markets and economics, politics, global events, renewable energy and alternative resources, new resources and development of new oil extraction technologies, social & environmental policies, and consumption patterns may influence the crude oil market dynamics. Such impacts and resultant price fluctuations might be very complex and may occur at different frequencies.

Classical econometric models such as random walk, autoregressive integrated moving average (ARIMA), error correction model (ECM), generalized autoregressive conditional heteroscedasticity (GARCH) model are used for crude oil price prediction. Recently, machine learning (ML) methods such as artificial neural network (ANN) and support vector machine (SVM) are used for the crude oil price prediction, which provide powerful tools to model nonlinear behavior or crude oil market dynamics (Jammazi and Aloui 2012; Lanza, Manera, and Giovannini

2005; Hou and Suardi 2012; Basiri 2015; Yu, Zhao, and Tang 2017; Murat and Tokat 2009; Kilian and Murphy 2014; Javadnejad 2012).

- *Objectives*

In this work, an ML model is proposed to predict crude oil price using multiple infulencing factors. The predictions are casted on multiple time-series to consider for complex factors that imact the market dynamics in differenct frequencies.

This report is structured as follows. Section 2.1 summarizes the datasets that are used in this study. Section 2.2 describes the data preparation and data wrangling procedures. In Section 2.3, the exploratory data analysis and feature engineering approcahes for ML training are described. Section 2.4 covers the ML pre-processing, training, models selection, model metrics. In Section 2.5, the final results for model training and validation are presented, as well as the predictions for 6-month time frequencies. We present the discussion of our results in Section 3. Finally, in Section 4, we summarise our recoomendaiton for future work.

2. Methodology

We use PyCaret, an open-source low code Python library that automates machine learning (AutoML) models to construct and deploy the models (Moez 2022). The library manages twenty-five different algorithms for regression, such as Extra Trees Regressor (ET), Gradient Boosting Regressor (GBR), Extreme Gradient Boosting Regressor (XGB), Random Forest Regressor (RF), Linear Regression (LR), AdaBoost Regressor (ADA), and eighteen other algorithms for classification.

We compare the performance of twenty-five AutoML models based on coefficient of determination (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Then, we select the top five models for each time-series and build ensemble models. Ensemble methods benefit different training algorithms for increasing the training accuracy for reaching a higher testing accuracy to substantially improve the accuracy of the integerated model (Ardabili, Mosavi, and Várkonyi-Kóczy 2020).

Finally, the multi-frequnecy prediction time-series are weight-averaged based on the perfromance of the ML model into a single integrated predication series that represent the final oil price predictions.

2.1. Datatets

The factors that influence the crude oil market dynamics include supply and demand, financial markets, politics, global events, alternative resources, development technologies, policies, and consumption patterns (Hamilton 2008; Hamilton 2009; Kilian and Murphy 2014; Zhao, Li, and Yu 2017; Lu et al. 2021; Wang, Wu, and Yang 2015). We use the crude oil prices of West Texas Intermediate (WTI) benchmark as the target feature. To take into account the aforementioned influencing factors a total of 32 feature variables were selected from publicly accessible data sources (EIA 2022a; FRED 2022; Investing 2022; WSJ 2022). Table 2 provides a list of the selected features, a description about each feature, and the sources of data.

Table 2. Selected dataset of feature variables for crude oil price

Category	Symbol	Variable	Unit	Source
Crude Oil Price	WTIPUUS	West Texas Intermediate Crude Oil Price	dollars per barrel	EIA
Supply	COPR_OPEC	Crude Oil Production, Total OPEC	million barrels per day	EIA
	PAPR_NONOPEC	Crude Oil Production, Total non-OPEC	million barrels per day	EIA
	INTL.55-1-WORL-TBPD	Crude Oil Production, NGPL, and other liquids production, World	thousand barrels per day	EIA
	COPRPUS	Crude Oil Production, U.S.	million barrels per day	EIA
Replacement Cost	RNGWHHD	Henry Hub Natural Gas Spot Price	dollars per million btu	EIA
Demand	PATC_OECD	Liquid Fuels Consumption, Total OECD	million barrels per day	EIA
	PATC_NON_OECD	Liquid Fuels Consumption, Total non-OECD,	million barrels per day	EIA
	FEDFUNDS	Federal Funds Effective Rate	percent, not seasonally adjusted	FRED
	IGREA	Index of Global Real Economic Activity	index, not seasonally adjusted	FRED
	CICPIUS	US Consumer Price Index (CPI): All Commodities	index, 1982-1984=1.00	EIA
	USACPIENGMINMEI	US Consumer Price Index (CPI): Energy for the United States	index 2015=100, not seasonally adjusted	FRED
	WPCPIUS	US Producer Price Index (PPI): All Commodities	index, 1982=1.00	EIA
	WP57IUS	US Producer Price Index (PPI): Petroleum	index, 1982=1.00	EIA
	EA19PIEAMI01GPM	roducer Price Index (PPI) of Euro Area (19 Countries)	index 2015=100, not seasonally adjusted	FRED
	ZOMNIUS	US Manufacturing Production Index (PMI)	index, 2017=100 (seasonally adjusted)	EIA
Inventory	PASC_OECD_T3	Petroleum Inventory, Total OECD	million barrels, end-of-period	EIA
	PASXPUS	Petroleum Inventory, US Total	million barrels, end-of-period	EIA
	COSQPUS	US Crude Oil Inventory: Strategic Petroleum Reserve (SPR)	million barrels, end-of-period	EIA
	COSXPUS	US Crude Oil Inventory: Non-SPR	million barrels, end-of-period	EIA
Monetary Market	RTWEXBG	Real Broad Dollar Index	index Jan 2006=100, not seasonally adjusted	FRED
	DXY	US Dollar Index (DXY)	index	Investing
	DEXUSEU	U.S. Dollars to Euro Spot Exchange Rate (DEXUSEU)	US dollars to one euro, not seasonally adjusted	FRED
Stock Market	SPX	S&P 500 Index	index	WSJ
	DJI	Dow Jones Industrial Index	index	WSJ
	COMP	NASDAQ index	index	WSJ
Commodity Market	Gold_Future	Gold Futures Historical Data	dollar per ounce	Investing

	Copper_Future	Copper Futures Historical Data	dollar per pound	Investing
Policy Uncertainty	GEPUCURRENT	Global Economic Policy Uncertainty Index: Current Price Adjusted GDP	index, not seasonally adjusted	FRED
Technology	MGWHUUS	Refiner Wholesale Gasoline Price	cents per gallon	EIA
	DSWHUUS	Diesel Fuel Refiner Wholesale Price	cents per gallon	EIA
	BREPUUS	Brent Crude Oil Spot Price	dollars per barrel	EIA

2.2. Data Cleaning and Data Wrangling

We used Jupyter Notebook 6.5.2 (Kluyver et al. 2016) and Python 3.9.15 (Python Software Foundation 2022) to process the data. The features in Table 2 were read through APIs (if available) or were downloaded directly from the data source. The data were initially set to be imported with monthly intervals or averaged to monthly values, then were limited to the target time frame of January 2000 and December 2022. The features were index based on their date values and then all merged together on the date values to create corresponding feature values for each month. Figure 1 shows the monthly West Texas Intermediate (WTI) crude oil price in the target time frame.

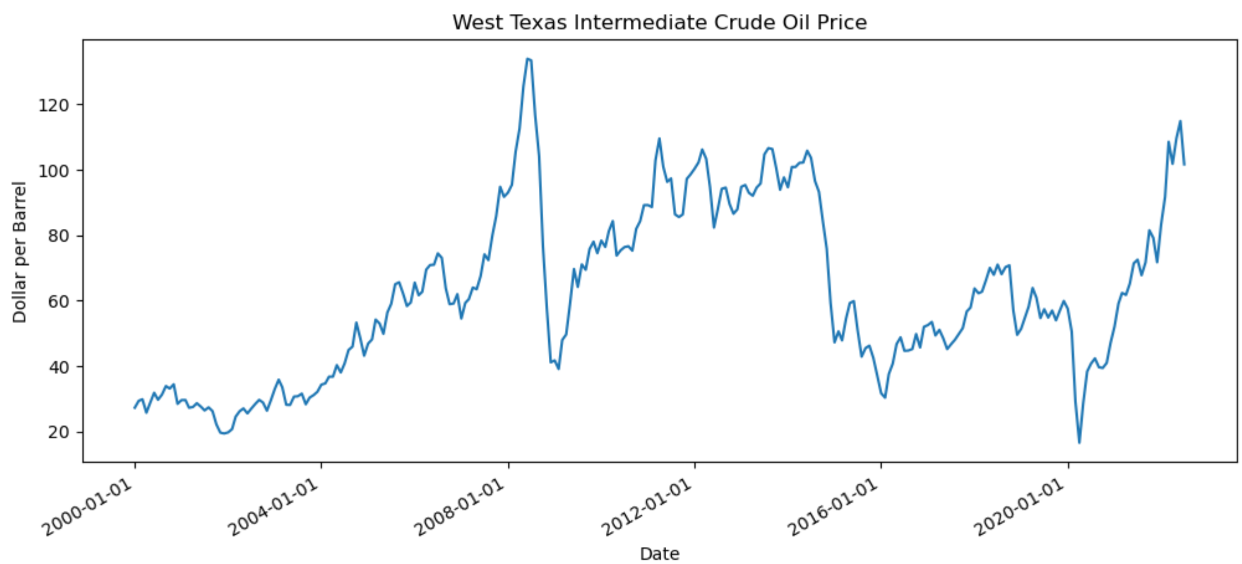


Figure 1. West Texas Intermediate (WTI) crude oil price

We used Pandas 1.5.2 (McKinney 2011) that is an open-source, simple, powerful, and flexible library for data analysis and data manipulation, Pandas is for Python programming language. The final dataset sized (274, 32) that represent 274-month records (rows) for 32 feature variables (columns).

An important step of data wrangling is dealing with missing data. Table 3 shows the summary statistics of missing data in feature variables. Missingno (Bilogur 2018) is also useful tool that provides a series of visualisations for presence and distribution of the missing data within a pandas dataframe. Figure 2 visually shows the distribution of the missing. To handle the missing data, features with more than the 10% of missing data were dropped from the dataset. The columns that had less than 1% missing features were imputed by using back and forward fill methods. For the remaining missing data between 1% and 10%, the rows for all features were dropped to create a dataset with no missing data. After treating missing data, the final dataframe sized (271, 30).

Table 3. Summary statistics of missing data in feature variables

Variable	Count	Percentage
oil_production_world	3	1.1%
petroleum_inventory_oecd	36	13.1%
global_real_econ_activity_index	1	0.4%
us_cpi_energy	1	0.4%
eu19_ppi	2	0.7%
real_dollar_index	72	26.3%
global_econ_policy_uncert_index	1	0.4%

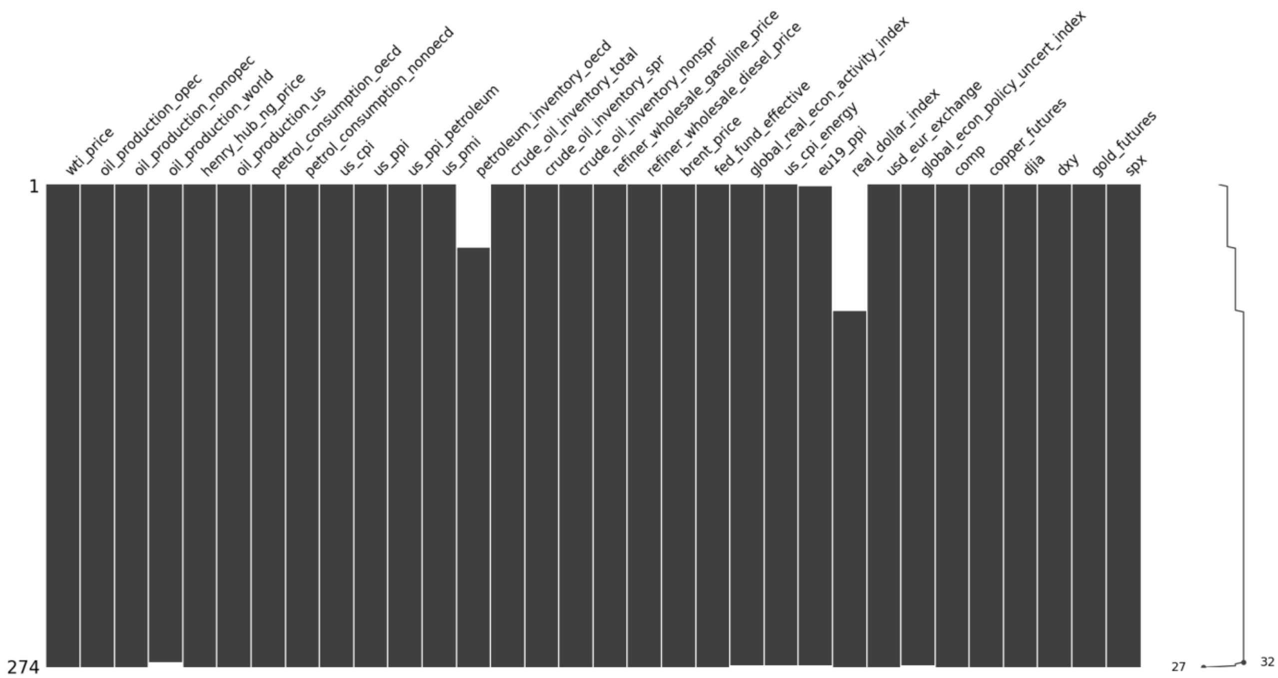
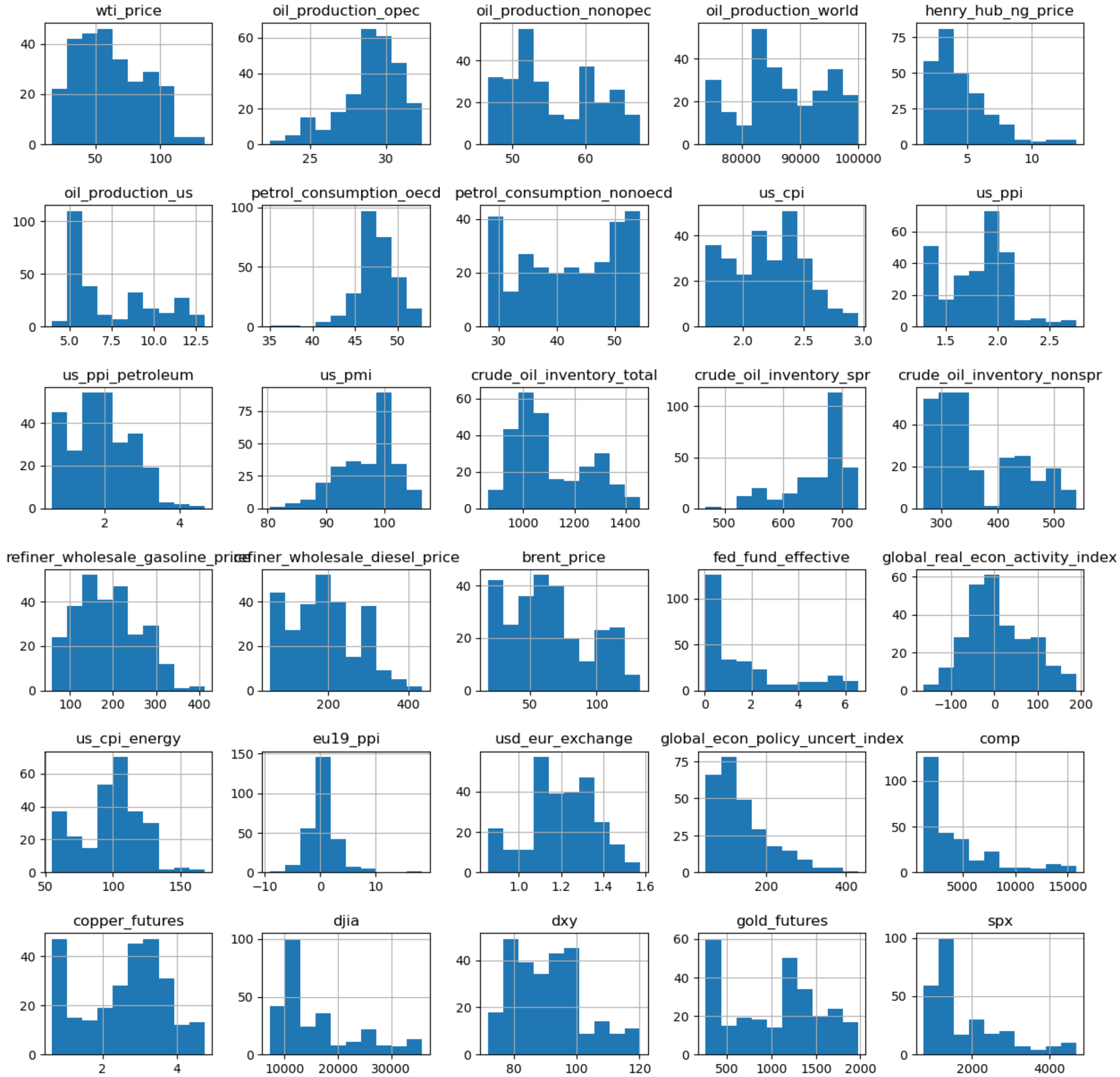


Figure 2. Missing data plot



2.1. Exploratory Data Analysis and Feature Engineering

2.2. Pre-processing and Training

2.3. Modelling

3. Results and Discussion

4. Recommendations and Future Work

- Fused TIR and RGB 3D models generated from UAS imagery offer great potential for mapping heat loss, supplementing non-destructive testing of structures, aiding in the inspection of electrical parts, and more.
- This study tested a simplified approach for generating 3D TIR point clouds from coacquired TIR and RGB images for remote sensing applications. The constructed TIR point clouds are georeferenced to the same coordinate system as the RGB clouds. The resultant point cloud preserves the spatial density and resolution of the RGB point cloud while adding TIR attributes.
- The integrated visualization approach tested in this study enables 3D point cloud and 2D raster representation of RGB and TIR data in one model, enhancing the visual interpretation and analysis of the remotely-sensed data.
- The approach does not require additional depth sensors, such as lidar, or GNSS-aided INS for registration purposes.
- In general, the approach is appropriate for cases when….. For evaluation, and as examples of implementation…. While the SfM processing of RGB images was able to generate reliable….

- In future work, the proposed integration and visualization can be integrated into standard …... Radiometric calibration was considered beyond the scope of the present study; however, in-situ radiometric calibration of the thermal camera might improve the spectral content of the data. As an alternative …...
- TIR-RGB image feature matching and auto-registration can handle non-synchronized dual-head camera captures; however, extraction of identical features and co-registration based on the extracted pair is challenging for images of different spectral bands at the scene without well-designed calibration patterns.
- It is recommended that follow-on studies be conducted to address these topics…..

Acknowledgments

We thank … for their valuable comments and suggestions on improving the quality of this paper.

We are thankful to …. for their help with collecting the data,

We would like to acknowledge … for supplying the logistics for

We also appreciate …. for providing surveying equipment and/or software.

We would like to thank …. anonymous reviewers for their constructive suggestions and comments.

References

Ardabili, Sina, Amir Mosavi, and Annamária R. Várkonyi-Kóczy. 2020. "Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods." In , 215–227. doi:10.1007/978-3-030-36841-8_21.

Basiri, Mohammad Hossein; Javadnejad, Farshid; Saeedi, Azita. 2015. "Forecasting Crude Oil Price with an Artificial Neural Network Model Based on a Regular Pattern for Selecting of the Training and Testing Sets Using Dynamic Command-Line Functions." In *24th International Mining Congress and Exhibition of Turkey-IMCET'15*.

Bilogur, Aleksey. 2018. "Missingno: A Missing Data Visualization Suite." *The Journal of Open Source Software* 3 (22). The Open Journal: 547. doi:10.21105/JOSS.00547.

Deng, Chao, Liang Ma, and Taishan Zeng. 2021. "Crude Oil Price Forecast Based on Deep Transfer Learning: Shanghai Crude Oil as an Example." *Sustainability* 13 (24): 13770. doi:10.3390/su132413770.

EIA. 2022a. "Opendata - U.S. Energy Information Administration (EIA)." https://www.eia.gov/opendata/.

EIA. 2022b. "International Energy Statistics, Total Oil Production." *U.S. Energy Information Administration*.

FRED. 2022. "St. Louis Fed Web Services: FRED® API." https://fred.stlouisfed.org/docs/api/fred/.

Hamilton, James. 2008. *Understanding Crude Oil Prices*. Cambridge, MA. doi:10.3386/w14492.

Hamilton, James. 2009. *Causes and Consequences of the Oil Shock of 2007-08*. Cambridge, MA. doi:10.3386/w15002.

Hou, Aijun, and Sandy Suardi. 2012. "A Nonparametric GARCH Model of Crude Oil Price Return Volatility." *Energy Economics* 34 (2): 618–626. doi:10.1016/j.eneco.2011.08.004.

Investing. 2022. "Investing.Com - Stock Market Quotes & Financial News." https://www.investing.com/.

Jammazi, Rania, and Chaker Aloui. 2012. "Crude Oil Price Forecasting: Experimental Evidence from Wavelet Decomposition and Neural Network Modeling." *Energy Economics* 34 (3): 828–841. doi:10.1016/j.eneco.2011.07.018.

Javadnejad, Farshid. 2012. "Presenting a Model for Prediction of Crude Oil Price Based on Artificial Intelligent Hybrid Methods and Time-Series." Tarbiat Modares University.

Kilian, Lutz, and Daniel P. Murphy. 2014. "THE ROLE OF INVENTORIES AND SPECULATIVE TRADING IN THE GLOBAL MARKET FOR CRUDE OIL." *Journal of Applied Econometrics* 29 (3): 454–478. doi:10.1002/jae.2322.

Kluyver, Thomas, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, et al. 2016. *Jupyter Notebooks: A Publishing Format for Reproducible Computational Workflows.* Vol. 2016. https://jupyter.org/.

Lanza, Alessandro, Matteo Manera, and Massimo Giovannini. 2005. "Modeling and Forecasting Cointegrated Relationships among Heavy Oil and Product Prices." *Energy Economics* 27 (6): 831–848. doi:10.1016/j.eneco.2005.07.001.

Lu, Quanying, Shaolong Sun, Hongbo Duan, and Shouyang Wang. 2021. "Analysis and Forecasting of Crude Oil Price Based on the Variable Selection-LSTM Integrated Model." *Energy Informatics* 4 (S2): 47. doi:10.1186/s42162-021-00166-4.

McKinney, Wes. 2011. "Pandas: A Foundational Python Library for Data Analysis and Statistics." *Python for High Performance and Scientific Computing* 14 (9). Seattle: 1–9. https://pandas.pydata.org/.

Moez, Ali. 2022. "PyCaret: An Open Source, Low-Code Machine Learning Library in Python."

Murat, Atilim, and Ekin Tokat. 2009. "Forecasting Oil Price Movements with Crack Spread Futures." *Energy Economics* 31 (1): 85–90. doi:10.1016/j.eneco.2008.07.008.

Python Software Foundation. 2022. "Python." Python Software Foundation (PSF). https://www.python.org/.

Wang, Yudong, Chongfeng Wu, and Li Yang. 2015. “Forecasting the Real Prices of Crude Oil: A Dynamic Model Averaging Approach.” *SSRN Electronic Journal*. doi:10.2139/ssrn.2590195.

WSJ. 2022. “The Wall Street Journal - Breaking News, Business, Financial & Economic News, World News and Video.” https://www.wsj.com/.

Yu, Lean, Yang Zhao, and Ling Tang. 2017. “Ensemble Forecasting for Complex Time Series Using Sparse Representation and Neural Networks.” *Journal of Forecasting* 36 (2): 122–138. doi:10.1002/for.2418.

Zhao, Yang, Jianping Li, and Lean Yu. 2017. “A Deep Learning Ensemble Approach for Crude Oil Price Forecasting.” *Energy Economics* 66 (August): 9–16. doi:10.1016/j.eneco.2017.05.023.