

# WTI Crude Oil Price Forecasting for the Next Year Using ARIMA Model

Tianxiang Wang\*

University of California, San Diego, United States

\*Corresponding author: tiw014@ucsd.edu

**Abstract.** The analysis and prediction of crude oil prices carry significant importance due to the increasing prominence of crude oil in the global economic market and its impact on individuals' daily lives. The crude oil price has significance in other industries, and there is an increasing presence of derivatives such as stocks and options. An escalation in the price of crude oil may result in an upward trend in gasoline prices, augmented shipping expenses, and amplified input costs for companies. This study utilizes the WTI crude oil price data from the first day of each month from January 2019 to September 2023 to provide a projection for the WTI crude oil price in 2024. During the investigation, identifying a trend demonstrating normalcy within the dataset has been utilized to effectively incorporate the ARIMA model, thereby serving the objective of forecasting. The phenomenon of bounce between up and down experiences a transition towards a flattened pattern by September 2023. Based on our forecasting analysis, it has been determined that the crude oil price is projected to exhibit a sustained plateau in 2024. While minor fluctuations in price may occur, the overall trend indicates a steady value of approximately 100 USD/Bbl.

**Keywords:** WTI crude oil; ARIMA model; forecasting.

## 1. Introduction

Most economists agree that crude oil is a major factor in economic expansion around the world. Estimates place oil and gas drilling's contribution to global GDP at around 3%, and the US imports 40% of its petroleum products [1]. It is important to emphasize, in light of the substantial part that crude oil plays in the global market, that the price of this commodity has an increasingly significant impact on the economic dynamics of a variety of other sectors. In addition to this, the occurrence of this phenomena has resulted in the need for a growing number of derivative goods [2]. The trading of stock and options related to crude oil has gained momentum, resulting in increased economic relevance of crude oil prices. The relationship between oil prices and economic activity may be understood by analyzing the standard supply-side effects. This implies that a rise in oil prices has the effect of elevating production expenses and diminishing the availability of the key input for production, hence hindering the rate of activity expansion and diminishing productivity [3]. Furthermore, according to a renowned essay by James Hamilton in 1983, it is posited that with the exception of one instance, every oil price escalation has been accompanied by a subsequent economic downturn in the United States since the conclusion of World War II [4]. Forecasting the price of crude oil holds substantial significance from the perspectives of economic market and productivity. The West Texas Intermediate standard is well recognized as a very active trading commodity on a global scale, making it a significant and instructive asset for study.

This paper aims to analyze the forecasted outcome of the West Texas Intermediate (WTI) crude oil price in USD in the year 2024. The data utilized for this analysis spans from January 2019 to September 2023, obtained from the reputable source, Trading Economics. This paper will consist of three main sections: the analysis procedure, the ARIMA model methodology, and the conclusion and discussion of the findings.

## 2. Method

### 2.1. Data

The primary dataset included in this research comprises the prevailing WTI crude oil price at 12:00 AM on the initial day of every month spanning from March 1983 to September 2023. In order to enhance model compatibility and optimize computational efficiency, this study focuses exclusively on monthly data spanning the most recent five-year period from January 2019 to September 2023. The objective is to forecast the total pricing trajectory for the year 2024. As a result of the limitations in data frequency, this study will provide a forecast at a macroscopic level, which may not offer daily or weekly precision.

### 2.2. ARIMA model

The present study used the classic ARIMA and auto ARIMA models to create the forecasting outcome. The Autoregressive Integrated Moving Average (ARIMA) model is commonly represented in the generic form  $ARIMA(p, d, q)$ , where the symbol "p" is used to indicate the ordering of the auto-regressive component, "d" is used to describe the order of differences, and "q" is utilized to signify the order of the moving average process. The model under consideration has both seasonal and non-seasonal components and is often denoted as  $(P, D, Q)$ . Here, P, D, and Q indicate the seasonal aspects of the model, while S specifies the number of periods each season [5]. The conventional ARIMA model requires manual adjustment of parameters in order to identify the optimal values that yield the most accurate results (minimize sic). These results are then evaluated using RMSE, MSE, MAE, and MAPE metrics. The conventional approach typically entails a significant time investment due to the need for manual adjustments. The ideal solution may go undetected if one assumes it has already been identified. In order to mitigate the potential for mistakes in determining the ideal parameters for the ARIMA model, this work employs the autoARIMA approach, which autonomously conducts a stepwise search to identify the optimal parameters.

The Mean Square Error (MSE) is a commonly employed metric in regression models, mainly when the independent variable represents continuous target values. The metric is quantified as the average of the squared discrepancies between the observed output and the estimated production [6]. The MSE is often favored due to its faster computational speed than the RMSE. However, it should be noted that MSE squares the error, which presents a limitation. Consequently, this paper acknowledges that relying solely on MSE to measure model accuracy is inadequate. Instead, a combination of MSE, RMSE, MAE, and MAPE is proposed to provide a more comprehensive assessment of the model's accuracy.

The root mean square error (RMSE) is a metric closely related to the MSE, with the key distinction being the inclusion of a square root operation applied to the MSE. RMSE exhibits similar characteristics to that of the standard deviation of the residuals, serving as an indicator of the dispersion of the residuals [6]. Not all errors contribute equally to the RMSE. Instead, it assigns greater importance to the most substantial error, allowing RMSE to assess the presence of a significant error during model training.

The Mean Absolute Error (MAE) cost function is calculated by averaging the absolute differences between the actual and anticipated output [6]. MAE is accessible from the impact of outliers because it does not involve squaring the residuals. The Mean Absolute Percentage Error (MAPE) is calculated as the average of the absolute values of the percentage errors (APE) [7]. Nevertheless, the MAPE exhibits a notable drawback: it yields uncertain or infinite results when the actual numbers approach 0 or are precisely zero. This constraint is often seen in specific fields [7]. The code has been modified in the present investigation to remove numerical values equal to zero or near zero. However, this modification may potentially impact the accuracy of the MAPE in this research. As previously stated, evaluating the model's correctness will encompass all four measures.

### 3. Analysis

The process of dataset analysis and forecasting has three primary components: data cleaning, model training, and to actually perform forecasting, which are fundamental aspects of academic research and analysis. Data cleaning involves the systematic identification and rectification of errors, inconsistencies, and missing values within a dataset. Model training refers to the process of developing and refining a statistical or machine learning model.

In the framework of data cleansing, this study initially investigates the presence of missing values or discrepancies within the dataset. Subsequently, the study proceeds to visualize the monthly crude oil prices (Fig.1). The dataset under consideration has been thoroughly recorded, confirming the lack of any missing values within it. Furthermore, the study uses probability distribution as a means of visually representing the dataset. The information supplied, which spans a duration of five years and includes WTI crude oil prices, has characteristics that suggest a normal distribution (Fig.2). The present study utilizes the rolling mean and rolling standard deviation methods to evaluate the stationarity of the dataset being examined. The word "stationary" refers to a condition in which a process stays in a state of statistical equilibrium, distinguished by probabilistic features that do not manifest any temporal variations [8]. Based on the examination of the dataset's bouncing rolling mean, it can be deduced that the dataset lacks the characteristic of stationarity (Fig.3). The subsequent procedure involves the extraction of both the trend and seasonality components from the dataset, as these will be stationary to analyze [9]. It is evident that the dataset exhibits a fluctuating trend, wherein the price initially decreases, afterwards increases, has a slight decline, and eventually stabilizes throughout this five-year timeframe (Fig.4).

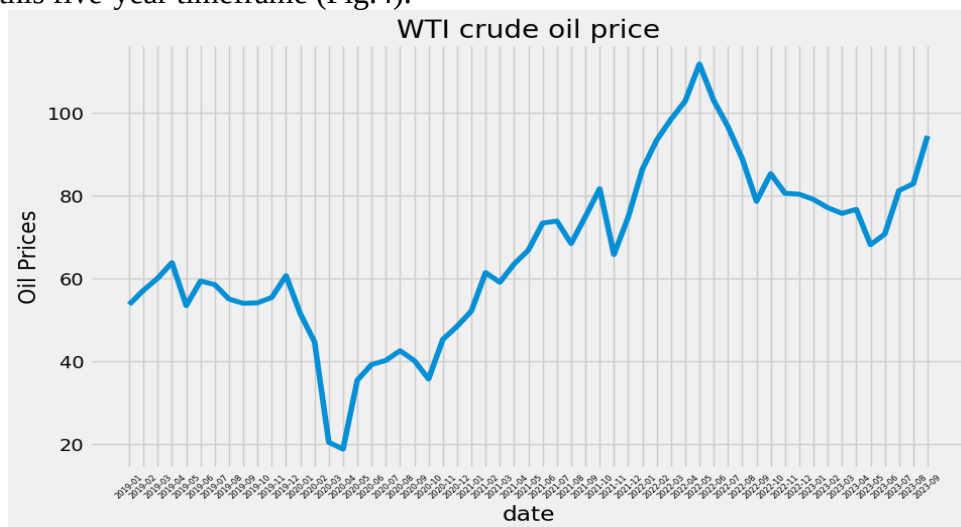


**Fig. 1** Visualization of the Data

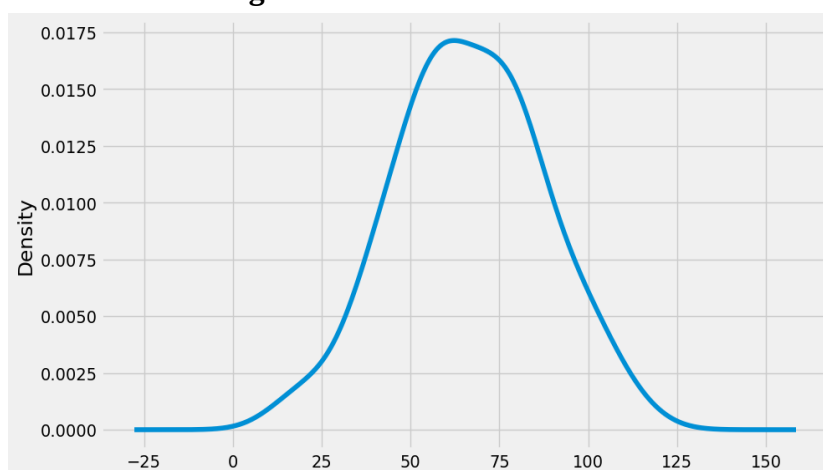


**Fig. 2** Density of the price

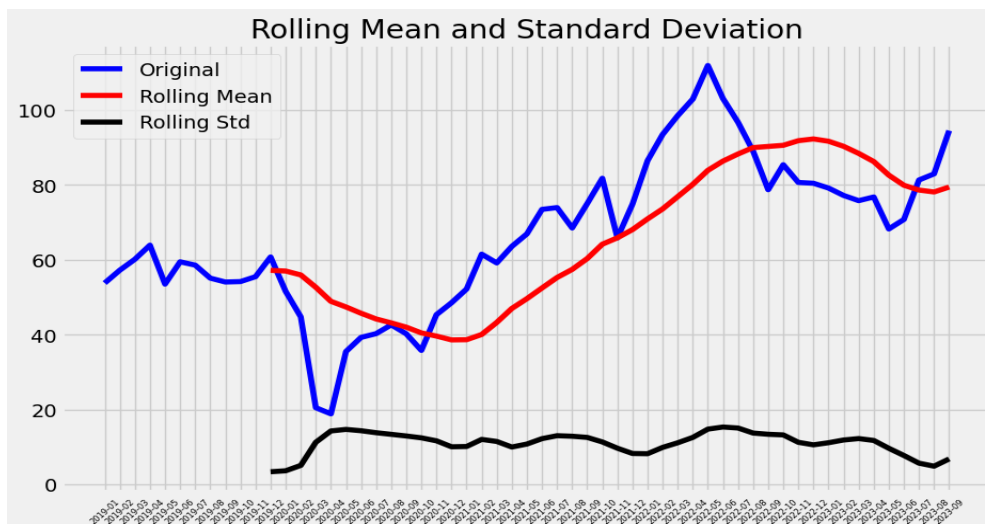


**Fig. 3** Rolling mean and rolling standard deviation of the price

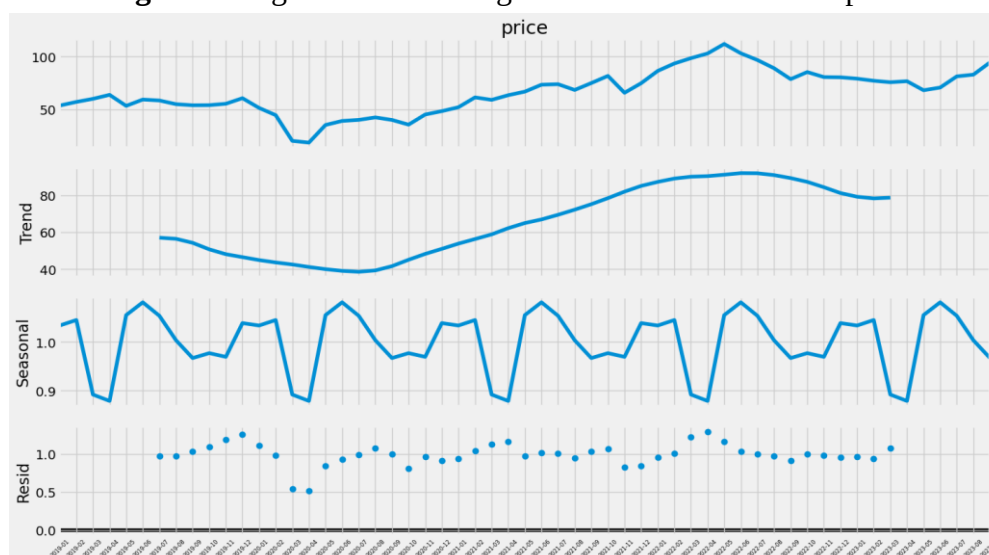


**Fig. 4** Trend and seasonality

Following the completion of data cleaning and visualization, the dataset including WTI crude oil price underwent a division into two subsets, with 70 percent allocated for training purposes and the remaining 30 percent designated for testing. Next, the training dataset is subjected to the autoARIMA procedure in order to determine the optimal parameters  $p$ ,  $q$ , and  $d$  for the subsequent model fitting stage. Once the ideal parameters have been chosen, the training set is fitted into an ARIMA model. The values of  $p$ ,  $q$ , and  $d$ , which were determined to be optimal in the previous phase, are manually specified for this purpose. Once the training set has been incorporated into the ARIMA model, the testing set is utilized to develop a forecast. This forecast is then employed to assess the correctness of the model (Fig.5). The MSE was computed to be approximately 0.04, while the MAE was found to be around 0.18. The RMSE was estimated to be approximately 0.2. However, the MAPE could not be determined due to the presence of zero values when calculating MAPE. Certain procedures have been implemented to remove the occurrences of zero values, resulting in a MAPE of 0.0. However, this adjustment may have an effect on the overall accuracy of the MAPE metric. In general, it appears that the ARIMA model has a high level of accuracy.

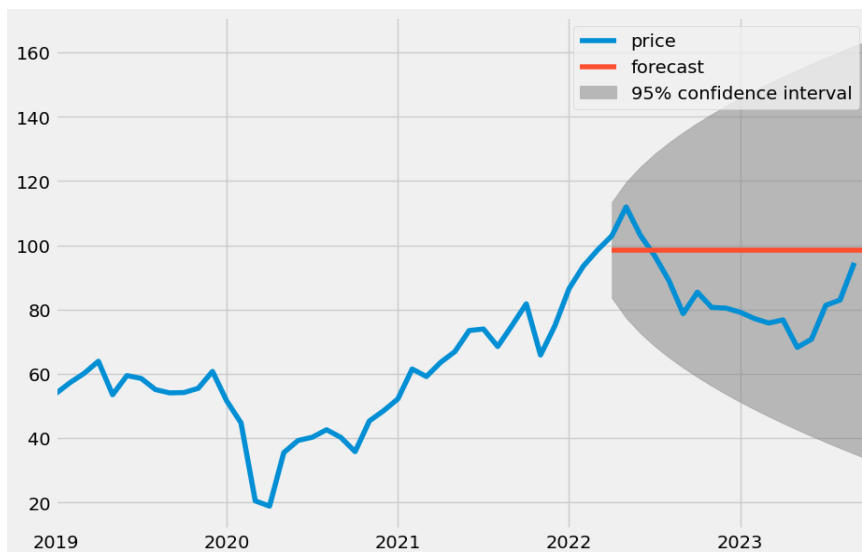


**Fig. 5** Forecasting of test data

## 4. Result

Utilizing the previously developed trained models, we can make predictions on the future trend of crude oil prices for the upcoming year, namely in the year 2024. The outcome is depicted in the following manner (Fig.6). Based on the forecasting results, it is evident that the crude oil price is projected to maintain a consistent trend which is around 100 USD/Bbl, similar to the previously observed flat pattern (Fig.4), when considering the visualization of both trend and seasonality for the upcoming year. However, the increasing confidence interval (CI) indicates a greater likelihood of price variations over time.

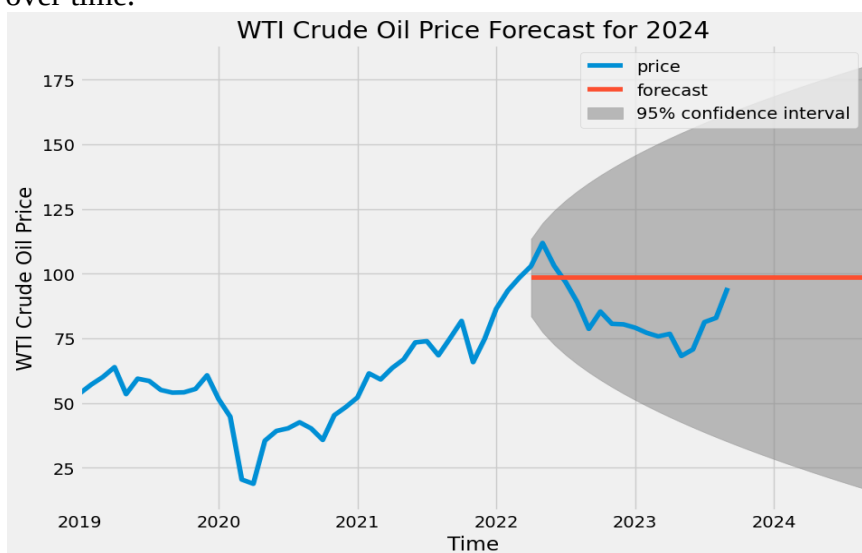


**Fig. 6** Price Forecast for 2024

## 5. Discussion

Despite the fulfillment of the normality assumption in our data, the ARIMA model possesses certain limitations in accurately forecasting turning points. In financial time series data, it is common to detect asymmetries, sporadic outbreaks occurring at irregular time intervals, and periods characterized by varying levels of volatility [10]. One of the primary characteristics of integrated financial models is the assumption of constant variance. However, it is worth noting that many financial data sets demonstrate fluctuations in volatility, which cannot be accommodated by this

assumption [10]. Furthermore, the data collected is on a monthly basis, indicating that forecasting based on monthly data may not accurately predict daily or weekly outcomes.

## 6. Conclusion

In summary, this study centers on the examination of the data pertaining to the WTI crude oil price during the past five years. Subsequently, utilizing this dataset, a projection is made regarding the anticipated price for the upcoming year, which is 2024. The forecasting process encompasses several key steps, including data cleansing, trend discovery, model fitting, and the actual forecasting. This study has chosen to employ the ARIMA model for the purpose of forecasting price trends in the forthcoming year. The findings indicate that the WTI oil price is projected to stabilize to around 100 USD/Bbl in the upcoming year at a macro level, but the widening confidence interval suggests the potential for minor fluctuations in both upward and downward directions. However, it is plausible to assert that the occurrence of substantial variations and shocks in the WTI oil price is improbable for the year 2024.

## References

- [1] Hong Miao, Sanjay Ramchander, Tianyang Wang, et al. Influential factors in crude oil price forecasting. *Energy Economics*, 2017, 68: 77-88.
- [2] Vesa Soini, Sindre Lorentzen. Option prices and implied volatility in the crude oil market, *Energy Economics*, 2019, 83: 515-539.
- [3] Farzana Alamgir, Sakib Bin Amin, The nexus between oil price and stock market: Evidence from South Asia, *Energy Economics*, 2021, 7: 693-703.
- [4] James, D. Hamilton. Oil and the macroeconomy since world war II. *Journal of Political Economy*, 1983, 91: 228-248.
- [5] Kumar Biswajit Debnath, Monjur Mourshed. Forecasting methods in energy planning models. *Renewable and Sustainable Energy Reviews*, 2018, 88: 297-325.
- [6] Kanishka Tyagi, Chinmay Rane, Harshvardhan, et al. Artificial Intelligence and Machine Learning. *EDGE Computing*, 2022, 53-63.
- [7] Sungil Kim, Heeyoung Kim. A new metric of absolute percentage error for intermittent demand forecasts. 2016, 32: 669-679.
- [8] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, et al. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Inc., 2016, 7.
- [9] Rob J Hyndman, George Athanasopoulos. *Forecasting: principles and practice*. OTEXTS, 2018, 3.
- [10] Andreea-Cristina Petrică, Stelian Stancu, Alexandru Tindeche. Limitation of ARIMA models in financial and monetary economics. *Theoretical and Applied Economics*, 2016, XXIII.4: 19-42.