CRUDE OIL PRICE ANALYSIS

Dissertation

DECEMBER 8, 2022

Student Name: Mausam Sudhir Jadhav
Student ID: 10605724
Coursework: Masters in Business Analytics

Abstract

The crude oil market has just entered a new phase of growth and the primary variables affecting crude oil have also changed. To determine and forecast the key influencing components, a framework is proposed in this research work. To understand how past events impact future occurrences, time series-based forecasting is vital. The performance accuracy of several time series models for oil prices is compared in this study. ARIMA, VAR, and LSTM models are explored in this research for the oil price prediction. . The VAR model is created for predicting the price. The dickey-fuller test is used to determine if the data is stationary or non-stationary. In both the testing and training sets the variable choice machine learning integration approach suggested in this research work outperformed the model without the core factor extraction and Univariate model. The results demonstrated that the proposed model's prediction accuracy is much higher than that of the Univariate model. The number of the key variables chosen by the BMA is neither the greatest nor the smallest, which shows that the number of core variables also influences the prediction outcomes. The comparative analysis showed that given a prediction with a 95% confidence interval, the proposed model outperformed the existing models. The VAR predictions often beat the ARIMA and random walk forecasts and have the least average forecast residuals and the best accuracy. This suggests using the random walk or ARIMA with caution when making forecasts, especially for short-term projections. The policymakers in the business should find the optimal forecasting approach using this comparison.

Table of Contents

List of figures

Chapter 1: Introduction

Crude oil referred to a petroleum-based product that is produced in the crust of the earth. Crude oil is a non-renewable and finite resource. This is one of the key sources of energy. This is refined into products such as diesel and gasoline that are useable petrochemicals. According to Wen, et.al (2019), Oil has become a basic need for human life. This is very essential for the World Transportation System (WTS) because with the help of oil they give energy for their vehicles to run. Crude oil is composed of hydrocarbon deposits and also from different organic materials which are formed by the remains of animals and this is used for many purposes like transportation fuels. Over more than millions of years, crude oil is formed from the remains of animals and plants which were crushed and heated by the layers of sand and rock until there is a large number of a pool of oil. The traditional way of measuring the oil volume is the barrel. This is important for the producers of the oil to understand and evaluate the value of the crude oil for each barrel. This is mostly traded for refined products.

If the price of crude oil will increase then it means that the price of the producing goods will also increase. The increasing price of crude oil will also affect costs just like it will increase the cost of transportation, heating, and also in manufacturing. This will be not good for the consumers because they have to give more money for oil. After all, the producer will increase the price by taking production costs from them (Aloui, et al., 2020). There are many disadvantages to the increasing price of crude oil, this will directly impact the currency of the particular country because the government has to invest more money in importing fuels. Some nations are dependent on furls to gather electricity, heat, and cooking also. The paper states that products made from petroleum and crude oil are held for almost one-third of the global consumption of earth. A few factors help to identify the effectiveness of OPEC in influencing oil prices which include the extent to which the members of OPEC can comply with the quotas of production. The stock market and crude oil prices are related to each other. The market of crude oil is not just limited to the standard methods of trading but the investors also bet on the prices of the crude oil using exchange-traded derivatives, mutual funds, and energy equities. *Multivariate and Univariate* time series-based forecasting methods are available in the market. In Univariate forecasting, the continuous variable that is output variable is predicted

using the previous data (Zhao, et al., 2015). This analysis necessitates the independent study of the outcomes for each pertinent dataset in the given dataset since it provides a quantitative statistical assessment. The Univariate analysis does not take into account any connection between the independent variables. A multivariate model is said to be less accurate for forecasting in comparison to the Univariate model.

This research work aims to discuss the key factors that affect the prices of WTI crude oil and conduct a critical analysis of the WTI crude oil market. The research will be helpful in the interpretation of the recent shift in the prices of crude oil to determine where it may be leading.

Problem statement:

The oil market has taken on new characteristics that have a substantial impact on the global economic development investor emotion and the country's strategic security. Meanwhile, as crude oil's financial qualities continuously improve the volatility of the prices is likely to affect the oil businesses' profitability and investor behavior (Su, et. al., 2021). As a result, a thorough examination of the features of the complicated international oil markets as well as reliable detection of new trends in crude oil prices are crucial.

Aim: This research aims to critically assess the recent change and develop a plausible prognosis for crude oil prices while considering the elements impacting it.

Objectives

- To set the hypothesis about the issue to prove the hypothesis
- Prove the hypothesis using inductive reasoning and factual interpretation
- Implement KDD approach for data mining to have the analysis of crude oil price
- To implement the linear regression model for determining the relationship between variables using the machine learning model for crude oil price analysis and prediction.
- Use the evaluation matrix for model evaluation.

Research question

RQ1: What arethe factors behind the change in WTI crude oil prices?

RQ2: What is the significance of machine learning in crude oil price analysis?

RQ3: What is the most effective machine learning model for the prediction and analysis of crude oil prices?

Structure of the research work

This research work is divided into multiple chapters. Chapter one provides the introduction to the research work, covering the aim, objectives, problem statement, and research questions. Chapter 2 provides the literature review that covers the analysis of multiple existing research papers to identify the resource gap. Chapter third presents the methodology which is used in this work. Chapter 4 provides the details of the data set along with the process of data set preparation and visualization. Chapter 5 presents the implementation of multiple learning models including linear regression and machine learning algorithms. Chapter 6 presents the evaluation and results using the evaluation matrix. Chapter 7 presents the conclusion of the complete research work along with the future research direction

Chapter 2: Literature review

According to Peng, et.al (2020), oil prices are set by global supply and demand, instead of the domestic production of any specific country. The top 3 largest oil producers in the world are the United States, Saudi Arabia, and Russia. There was an oil crisis in 1973, and then this was very important to make a new economic model of the global oil market, so the Research and Development Corporation which is an American nonprofit global policy launched this model in 1974. As per the view of Khan, et.al (2021), there were four sectors in this model they were production of crude, refining and transportation, these sectors were launched in different regions like the $United\ States, Canada, Latin\ America, Europe, the\ Middle\ East$, and $Asia$. The shortage of crude oil will increase the price of other elements like petrol, diesel, kerosene, etc. This will impact the world economy in a very negative term. In modern agriculture techniques, the supply of petroleum and natural gas is the most important because this will directly impact the food price. Inflation in oil prices is very dangerous for energy security and threatens food security because energy is required to supply food. So consumers will have to pay more taxes for their food because producers will take delivery costs also. In the future, there is a high chance of increasing the price of crude oil because it is in demand and there are not enough suppliers who can fulfill the requirements of people. According to Orojo, et.al (2019), there is more demand for

crude oil and not enough suppliers so this is a prediction that the price of crude oil will upsurge in the upcoming years.

But this oil is not yet ready for use so that's why it is known as crude oil. As per the review of Abdollahi, et.al (2020), when crude oil is ready for use then it can be used for different purposes. The main use of crude oil is for transportation. Many chemicals are prepared from crude oil when the refineries turn crude oil into usable products just as gasoline, diesel, aviation fuels, etc. Chemical industries need crude oil as raw material so that they can produce useful chemicals for the industry's uses, also crude oil is in demand by women because they use it for cosmetic purposes.

According to Lu, et.al (2020), there are 4 main types of crude oil: *light distillates*, *middledistillates*, *medium oil*, and *heavy fuel oils*. In light distillates oil, there are several varieties of petroleum, jet fuels, kerosene, gasoline, etc. Major of grade 1 and grade 2 oils and diesel is the part of middle distillate. In heavy fuels, there are grade 3,4,5,6 fuel oils, and heavy marine oils. All of these are very important for transportation. At the current time, crude oil plays the lead role in economic activities. Since the 1970s crude oil prices fluctuates. The price of crude oil and economic activities has a direct relationship with each other. As per the review of Li, et.al (2019), if the price of crude oil will increase then it means that the price of the producing goods will also increase. The increasing price of crude oil will also affect costs just like it will increase the cost of transportation, heating, and also in manufacturing. This will be not good for the consumers because they have to give more money for oil. After all, the producer will increase the price by taking production costs from them. There are many disadvantages to the increasing price of crude oil, this will directly impact the currency of the particular country because the government has to invest more money in importing fuels. Crude oil is very important for human life. The rising oil price can be seen everywhere not only in gas stations. According to Gupta, et.al (2020), oil is very important for today's life because it is a source of energy and is used in transportation, so because of the increasing price of oil the other prices of goods and services that we use, will also increase.

According to Su, et.al (2021), there are also some advantages to the increasing price of crude oil. This will be very beneficial for the environment. Refineries use lots of chemicals that are harmful to the environment and the waste of the chemicals and other products will go into the

water and atmosphere which can easily pollute it and destroy the lives of many fishes, animals, and other organisms. If the crude oil price will decrease then other products like petrol and diesel will also available at cheap rates. So, then everyone can easily purchase them without any money issues and then they will start to take unnecessary rides. As per the review of Sun, et.al (2022), The emission of these fuels are so harmful to the environment and if there will be a huge number of people who are using their vehicles then will plan the main role in polluting the environment. So, this is the main advantage of increasing the price of crude oil so that people will not drive unnecessarily. They will not want to waste their money, they will use their vehicles when it is necessary and this will be very beneficial for the environment. There is one more advantage of increasing the price of crude oil people will try to purchase electric vehicles, and they will easily attract to electric vehicles. According to Cen, et.al (2019), if the price of crude oil will increase then people want to save their money because it is useless to drive normal vehicles, they are already polluting the environment and taking lots of energy. So the rising in the price of crude oil is also good in many ways.

As per, Su, Huang, Qin, and Umar, 2021 the prices of crude oil are identified by supply and demand globally. The growth of the economy is a major factor that affects the product which is petroleum and also demand for crude oil is increasing. The growth of economies raises the demand for energy in normal and specifically to transport the product to the customers from the producers. The sector of transportation is highly reliant on products of petrol like gasoline and fuel. Some nations are dependent on furls to gather electricity, heat, and cooking also. The paper states that products made from petroleum and crude oil are held for almost one-third of the global consumption of earth. A few factors help to identify the effectiveness of OPEC in influencing oil prices which include the extent to which the members of OPEC can comply with the quotas of production. Also, the ability of the customers to decrease the consumption of petrol, and the competition amongst the producers when there is a change in the prices of crude oil. The researcher has also stated the difference between demand and supply of the oil market from the sources of non-OPEC and is also stated as the call on the members of OPEC as they have the role to maintain the overall production capacity of crude oil across the globe. Therefore, the paper of Su, Khan, Tao, and Nicoleta-Claudia, 2019 highlights the case of Saudi Arabia which is the leading oil producer of OPEC and also among the leading oil exporters in history. The development and maintenance of idle capacity of production are generally expensive to

companies dealing across national borders. Hence, the capacity of OPEC offers an indicator of the global oil market's capacity to reply to the current and potential changes in prices of crude oil in the world.

As per the research of Erickson, Lazarus, and Piggot, 2018 there are a few geopolitical events and differences that affect the supply of crude oil in West Texas and products of petrol to market that can affect the prices of crude oil and petrol products as well. The events might build unexpected situations for the future demand and supply of crude oil in West Texas that can affect the prices rapidly. Therefore, as per the researcher volatility in the prices of oil is tied to less response and less supply and request to the changes in rate within the short term period. Further, as per the analysis, the production of crude oil and the goods that need petrol products is a necessity for the energy sources and is fixed in the near time. As per, Hoang, et al., 2021 the government needs time to work on the development of new sources of supply or to differ the production, and when the prices increase, there is a shift towards other fuels. Hence, in such changing situations, there is a need for large changes in prices to have a balance in the supply and demand of products. Therefore, fluctuation in the supply and request for the rate brings a massive turn in the business.

Therefore, as per Mundia, Secchi, Akamani, and Wang, 2019 most of the reserves of crude oil across the globe are situated in a location that is more prone to the healing of political factors and in areas that might have had disruptions in the production of oil earlier due to political tasks. Further, several changes in the prices have shocked the people of West Texas and also at the same time affected the political balance. The disruptions of prices in West Texas have to lead to war and challenged the people. In the current year, issues and political operations taking place in West Texas have participated in the disruptions and high increase in the rate of crude oil in the market. As per past cases and events of disruptions in oil supply caused due to political events, the participants in the market have uniform access to the possibility of disruptions in the rate of crude oil shortly. Moreover, the paper stated that the size and period of the potential disruption and participants in the market are also considered as the availability of stocks of crude oil and different producers to deal with the loss in supply. Hence, in the case when there is less capacity then there is a massive effect of supply disruption on the prices that can be expected by considering the existing supply and demand. The paper has also highlighted the role of weather in the crude oil distribution across the globe as it can lead to a high rise in petroleum goods and a

decrease in the crude oil distribution from the Gulf to other countries. Moreover, in case of extreme cold, there can be a strain on the market because the producers of crude oil can attempt to supply accurate products of petroleum-like oil for heating to the customers in less time than placing orders.

The paper of Kocaarslan and Soytas, 2019 highlights that due to the price fluctuation of crude oil the market receives signals about the demand and distribution balance. Further, the high increase incline in prices highlights that there is a need for a high supply and in the situation of a price decrease there is a high supply for the existing demand. The future markets also offer data that is related to the physical supply balance and also the expectations of the market. Moreover, the major subject that is highlighted is that the outlook for the prices of crude oil is not certain and there can be any disruption in the same. Also, most of the regular day-to-day supplies in life come after the processing of crude oil which is existing in every phase of life. Therefore, any upward rise or downward rise in the prices of crude oil is a type of energy that cannot be properly generated and is often stated a massive effect on the market as well as the people. Moreover, in addition to this, crude oil is referred to as the energy type that is precious across the globe and has been gaining importance across the Brent, Tokyo Exchange, and London Stock Exchange also based on the Dubai rate for crude oil.

According to, Norouzi, de Rubens, Choupanpiesheh, and Enevoldsen, 2020 the data that has been gathered from the US Information Administration has shown that there is a high volatility of the rate of crude oil, and this further states that during the time frame of 1986 from 1985 there was a rise in the production till 5,255 million that also resulted to the decrease in the prices of the West Texas Intermediates. Further, later on, the crisis of crude oil during the year 1990 was because of the major Kuwait invasion. Therefore, the Kuwait invasion resulted in an increase in crude oil prices for the West Texas Intermediates and a decrease in the production rate. The paper has highlighted that because of this the economy of the US was in recession and in reply to this the Agency had activated the plan and forced the people of the West Texas Intermediates to put millions of dollars into the plan in a single day. As per the author in the year 2002, there was a depreciation in the rate of the dollar to 50% and during the major outbreak of the issues in the year 2007, the rate of the Dollar and Euro decreased at a massive rate. According to, Corbet, Goodell, and Günay, 2020 huge capital were absent from the market of the US and the markets were in huge debt. Therefore, at a uniform rate, the prices of crude oil increased and during the

year 2007, the rate of depreciation of the dollar was also increasing. A major understanding of the overall crisis was that there is a huge link between the rate of exchange and the rate of crude oil The paper has stated that post the rate of crude oil shifted in the West Texas Intermediates the record went high to 145 US Dollars and this triggered to crisis worldwide. The crisis led to a recession in the economy and huge decline in the Index of the US dollar and a decrease in the US Economy also.

As per, Obringer, Mukherjee, and Nateghi, 2020 in the West Texas Intermediates the prices of crude oil had a massive impact and lead to a recession in the economy as well. Further, the OPEC wished to have maintenance of crude oil price toll 80 dollars and needed the support of the OPEC members to apply a better cut on the facilities of production. The recession in economy of the West Texas was due to the decrease in the call out for crude oil which is more than the decrease in the distribution of crude oil. Moreover, considering the past data that of 31st December 2008, the West Texas Intermediates closed at less point. There was a reduction in the capacity of production to 520 thousand barrels in 2008. This also states that the government of many nations across the globe tried to protect the market by bringing better measures for the people. Also, the economy in 2011, the prices returned to an extreme level, and till 2014, the prices were crude oil also started to decline because of the rise in the capacity of production of crude oil in the USA and the decrease in its demand during the developing nations. Later, during the year 2015, the prices of crude oil in the West Texas Intermediates were decrease due to the reason that OPEC did not want to decrease the capacity of production to keep up with the prices. Rather than this, OPEC continued to increase the level of production, to guard its share in the market. Moreover, to deal with the crisis a proper alliance was formed involving Russia to bring a decrease in the capacity of production of crude oil, and this was done to bring better measures in the market and stabilize the economy.

As per the author, Johra, Heidelberg, and Le Dréau, 2019 the studies have impacted the capacity of producing crude oil of the countries having $OPEC\ and\ Non-OPEC$ members gathered from the period till 2012. The gathered data corresponded to the price of crude oil. The purpose of the investigation was to identify if the capacity of production of oil for the members of OPEC will have a major impact on the price of oil and if the members of non-OPEC will affect the production ability of the West Texas Intermediates. The outcome of the study has highlighted that during the phase of 1196 from 1974 there was a positive side of growth in the capacity of

crude oil production. Therefore, all the results have stated that the capacity of producing crude oil of the members of the OPEC has been amongst the most necessary factor that has affected the price of crude oil. The primary focus of the company was to ensure that there is effective management of crude oil for the OPEC members through the launching of the target. However, as per, Beckmann, Czudaj, and Arora, 2020 the author's study that during the past there was a reduction in the capacity production of crude oil that might have increased. As per the research, the outcome was just the same as the old and the study has highlighted that with the price increase fewer people invest in products related to petroleum which decreases the capacity of production. The author of the paper has also stated about owing to the decrease in the capacity of production of crude oil by the members of the OPEC and the rate of crude oil increased to 30 US DOLLARS. The rate of crude oil in the West Texas Intermediates went high as per the situation of the market and political events.

The author Aloui, Goutte, Guesmi, and Hchaichi, 2020 also stated the high rise in the rate of crude oil during the initial period of 2008 and the beginning was from 133.93 US dollars. Therefore, afterward, there was a decrease rapidly while till the end of December, the prices were shifted to the bottom. During the beginning period, the OPEC had observed but later the month of October and after witnessing a major decrease in the price of crude oil it decided to decrease exports of crude oil for the West Texas Intermediates. Also, the paper has mentioned that during the year, 2009 there would have a further decrease in crude oil export that was conducted to have proper maintenance of the crude oil prices. Moreover, it stated that the members of OPEC were able to control the prices of crude oil which would have impacted the overall economy due to the massive crisis. Hence, the research stated that with the price reduction the company could have made it to a huge outcome for the country in total. Therefore, this highlighted the effective connection between price and capacity of production of crude oil.

In the research paper, by Umar, Su, Rizvi, and Lobonț, 2021 the author used the autoregressive model in order to have a proper study of the relationship that is amongst the rate of exchanges of country and the prices of crude oil as well. Therefore, the results of the research highlighted that after prices of crude oil were leaked to the unauthorized people of the West Texas Intermediates, the shilling states about the necessities in three nations. However, during the event of making oil prices shock, the rate of exchange in the West Texas Intermediates could have a different response to the situation. Therefore, the studies were done with the aim to have to investigate the

closing rate of crude oil in the research phase of 2015. Further, the variable return of rate was put to use for proper analysis. Also, Gold highlighted that there was no positive correlation with the Index of the US Dollar and it highlighted the proper correlation with the prices of crude oil in the West Texas Intermediates and that was a lag of one term. Hence, in the West Texas Intermediates, there was negative relation with self-lag.

Solarin, 2020 made use of the ADF test to have a proper study for the period of research till 2016 in December and it was determined that prices of crude oil were not proper during that time. Therefore, all the variables had a proper GARCH model impact that was put to use to have a proper investigation about the bond amongst the rate of crude oil and the US dollar rate as well. Moreover, the rate for crude oil was affected and these variables have an impact on each other. Further, in this research, there was a couple of function used by the author to have a proper analysis of the recent shift in the West Texas Intermediates and where it might result in assessing the factors influencing the shift. Also, the main targets of the researcher were to investigate the rate of currency exchange and the amount of crude oil. The outcome states that there was a drop in the crude oil amount which lead to the increase in the Index of the US Dollar which had a negative effect on the currency and economy. The author also states that the stationary method was used to investigate the crude oil amount and dollar value along with the developing market. Kocaarslan, Soytas, and Soytas, 2020 researched the link between the volatility of crude oil and the rate of exchange. The period of research suggested different measures for the people of the West Texas Intermediates to deal with the crisis and have proper working. The dollar rate affected the prices of crude oil for the West Texas Intermediates and in other references; it could have been mentioned as the base for having investment. Also, the correlation states that the West Texas Intermediates had to deal with issues during the crisis and a major shift in prices affected the demand and supply for crude oil and products of petroleum.

As per, Yang, 2019 the crisis of high rise in the crude oil amount was during the phase of 2006 and 2004 and was a topic of huge debate. The debates include changes in the sector downstream and mainly the area of refining whereas some of the refineries in the country did not rise since the year 1981. Therefore, under such circumstances, the absence of refining capacity was seen as a reason for the increase in the amount of crude oil. The paper has focused on different factors that proposed to define the rapid increase in the capacity of producing crude oil and a non-linear form of relationship that is amongst the supply of oil products and prices of oil along it.

Moreover, the expectation stated the lack of long run that might affect the price of oil. The argument in the research stated the factors that were necessary for the factors of oil prices that have an impact on the capacity of production, and supply and demand of crude oil. Therefore, having casual links in the supply chain of the US highlights that the crude oil prices have been exogenous and the factors downstream like refinery rate of utilization did not have much effect on the crude oil prices. The trend growth was more challenging and the rules were fuzzy it is necessary to consider the historic period so that purpose to maintain the crude oil prices can be forecasted. Further, the historical data of crude oil had some benchmarks that offer references to the crude oil price. The main factor was to have an understanding of the West Texas Intermediates prices of crude oil and the link to evaluate the link of the amount of crude oil along with the production capability for the production of the crude oil and the impact it has on the market.

Research gap: The research gap is identified in this area from the literature review. It has been identified that the WTI crude oil prices are very volatile and such independent experts do not always give sufficient evidence as to why crude oil prices suddenly peak and plummet. Multiple underlying financial and technical elements influence the trade, but the specific source of the volatility cannot be predicted.

It might be due to a chain of reaction in the corporate structure in which one economic uncertainty causes a spike or decline in a certain commodity first stop search gaps might be identified in multilevel while examining the data

Chapter 3: Methodology

Existing data mining methodologies

CRISP-DM and KDD are the two common methodologies that are used in the data mining process. **CRISP-DM**: The CRISP-DM stands for $Cross\ IndustryStandard\ Processfor\ DataMining$. There are six steps included in this process i.e. understanding of the business, understanding of the data, the technique of modeling, evaluation, and deployment.

KDD: This stands for $Knowledge\ discovery\ in\ the\ databases$. This is the process of finding the importance of the data set. This is another way of mining the data. The different steps include preparation and selection, cleansing of the data, incorporation of the prior knowledge about the dataset, and interpretation of the specific and accurate solutions from the identified results (Aloui, et al., 2020).

Following are the tasks and subtasks that are following in this work under the methodology:

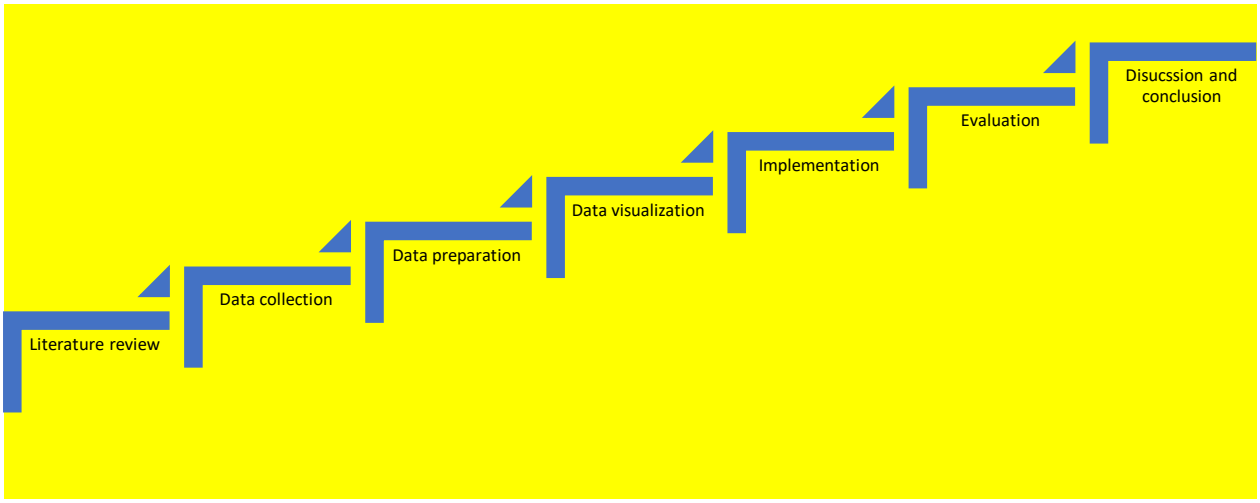


Figure 1 Methodology

1. **Literature review**

The papers are reviewed to get the qualitative information about the research area. Following sub-tasks are included in this step:

- Filter the papers by using keywords and time line

- Review the papers
- Collect the information and identify the research gap

2. **Data collection**

This step is based on the collection of the dataset for the analysis. The datasets are retrieved from different publically available data sources such as NYMEX and NADEX. Then the suitable variables are selected from these datasets to create the new dataset. Then the price is set as the target variables. The subtasks are listed below:

- Identify the sources for crude oil data
- Download the data from the sources
- Identify the appropriate features

3. **Data preparation**

The data is collected in the time-series format therefore multivariate analysis is performed using VARMAX. To perform the analysis, the abnormalities and outliers are identified in this work. The abnormalities are identified using the Pycarat module and interquartile range. Outliers are removed by removing the rows.

- Use the Interquartile range and Pycarat module
- Remove the outliers

4. **Data visualization**

The next step of the methodology is the data visualization. The visualization is done to explore the dataset and extract the useful information from it. For analyzing data, correlation analysis is performed. To visualize the results and extract the information that is useful in this work, graphs and charts are created. The Dickey-fuller test is used for determining that the data is stationary or non-stationary.

- Create the pivot table to filter the data
- Create the graphs and charts
- Conduct the dickey fuller test
- Perform the correlation analysis

5. **Implementation**

This stage of the methodology is based on implementing the time series modeling techniques for the analysis of data. Following are the sub-tasks of the implementation.

- Identify and compare the time series modeling technique
- Determine the model selection criteria
- Prepare the VAR model for the prediction of the oil price.

6. Evaluation

In this stage of methodology, the results are evaluated based on the evaluation criteria. Multiple types of tests are performed on the data for the analysis of VAR model. Statistical test is also conducted for the evaluation of the data. Following are the sub-tasks that are used in this work:

- Determine the evaluation criteria
- Perform the Granger's causality test
- Perform integration test
- Checking for stationary data using ADF test
- Serial correlation of the errors
- Forecasting the performance evaluation
- Inverting the transformation
- Statistics test

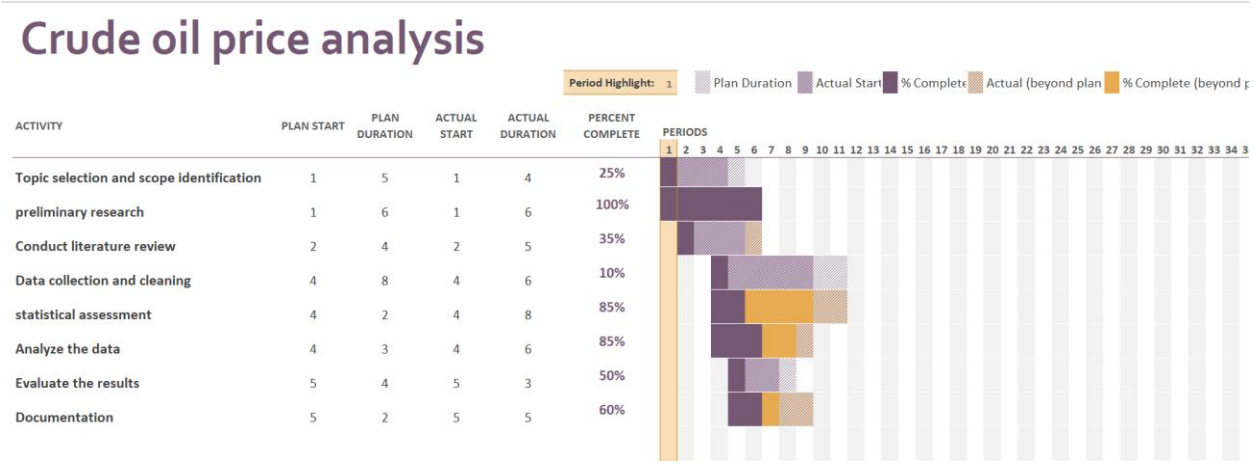
7. Discussion and conclusion

In this stage, the work done in this work is discussed and concluded. Following are the subtasks:

- Discuss the findings
- Conclude the work
- Present the future direction

Gantt Chart:

The management of the project is done using the Gantt chart. Following image shows the Gantt chart that is followed in the crude oil price analysis project. The key activities of the project include the selection of the topic, scope identification, preliminary research, conduct literature review, data collection and cleaning, analysis of the data, results evaluation and documentation. This shows the activities name, duration of the plan, actual start and actual duration. The timeline is showed in the form of weeks. This also shows the percentage completion of the project



Suitable approach

KDD approach is used in this approach as it is helpful to find new and possibly useful information from the massive volumes of data. It is helpful in determining the predictive nature of crude oil prices based on the patterns. This process is iterative which means that the information gained is fed back into the process to start improving its efficiency. As a result, the data gets better polished at each level. This establishes a loop that continues after the final outcome is implemented, flowing directly back into the goal-setting process.

Chapter 4: Dataset and visualization

The first step is to collect the data and get information about crude oil pricing. In order to build a good forecast view for the analysis of crude oil prices, the data is collected from previous years. There are publicly available data sets for WTI crude oil prices on Kaggle as well as commercial sites such as statistica.com. Such as NYMEX which is the New York Mercantile Exchange and NADEX are the most commonly used platforms for the process of crude oil Futures Trading. The prices of crude oil are traded in 1000 barrels for each and every contract. The live and historical data for crude oil prices are examined. The daily prices of crude oil are updated to the present closing day for a very long time on official statistics websites such as the US Energy Information Administration (Corbet, et al., 2020). The data set including the crude oil prices in CSV format can be derived from either of these sources. The data set is directly downloaded as it is readily available. However, this is not the handiest data set because most of such resources do not include essential elements impacting the prices of crude oil. As a result, it is necessary to identify individual characteristics that correspond to the changes in the prices of crude oil there are multiple elements that influence the prices of crude oil including demand, production, refinery utilization, global EV sales, REIF seasonal demand, and so on. These variables have a significant relationship with the prices of crude oil and can be used to create a new data set. In this dataset, the crude oil price is set as the object variable (Kocaarslan, and Soytas, 2019). The data set includes the daily closing crude oil prices from the year 2016 to the year 2022. There are 7 variables in the dataset apart from the crude oil price, including the date, utilization of the refinery in percentage, demand, production, contract 1FP, and global sale in millions, and last day's price.

Description of the dataset

The dataset includes the details about the crude oil data. The dataset is between the ranges of year 2016 to 2022. The dataset includes 8 columns that contain the data of date, utilization of refinery in %, production, demand, contract 1FP, global sales in millions, last Day's price.

	Utilization of Refinery in %	Production	Demand	Contract 1 FP	Global Sale (in millions)	Last Day's Price	Price
Utilization of Refinery in %	1						
Production	0.458794502	1					
Demand	0.475774788	0.981146912	1				
Contract 1 FP	0.441018426	0.498887197	0.433477	1			
Global Sale (in millions)	-0.600701035	-0.324257059	-0.41015	0.064245815	1		
Last Day's Price	0.441596776	0.49878633	0.433566	0.993051532	0.068556436	1	
Price	0.439125257	0.501449257	0.435813	0.99853653	0.07110847	0.993838393	1

Figure 2 Correlation analysis of the dataset

The correlation analysis allows exploring the dataset. The results of the analysis showed that the price of the crude oil is mainly related with the contract 1FP and last day price. The price of the crude oil is least affected by the Global sale of the crude oil. The global sale is also negatively correlated with the utilization of refinery, production and demand. From the exploration it can be determined that the global sale column can be removed from the data as it is not significantly related with most of the other variables.

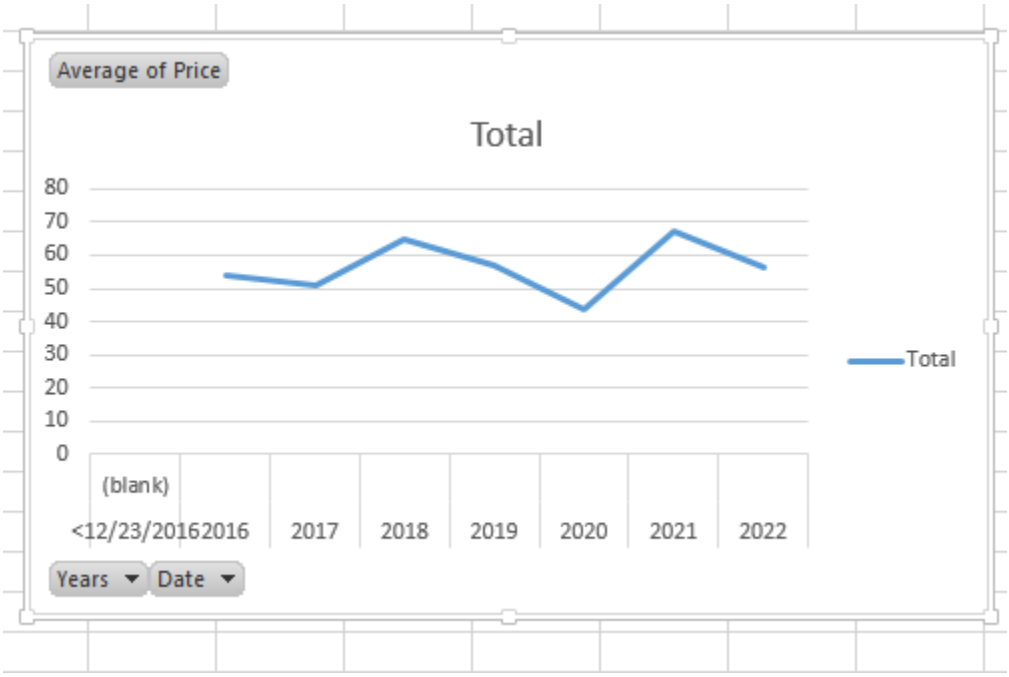


Figure 3 Price by year

The visualization of the data is showing that in year 2020, the price of the crude oil was lowest. It has been assumed that the decline is the price is caused by the COVID pandemic situation. The growth in the crude oil price is seen in the year 2021.

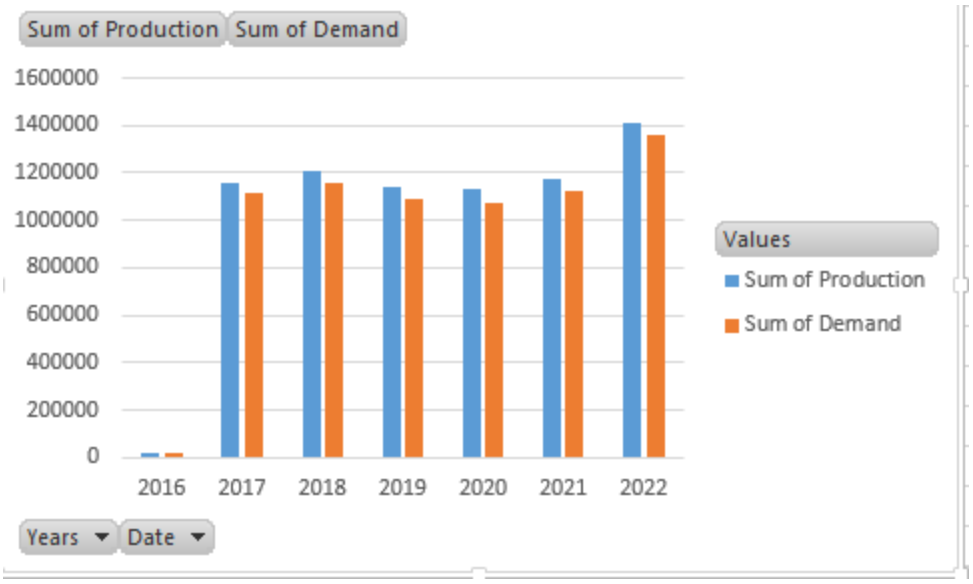


Figure 4 Production by year

The graph of the production by year is showing that the production amount for the crude oil is always greater than the demand throughout all the years. The maximum demand of crude oil is determined in the year 2022.

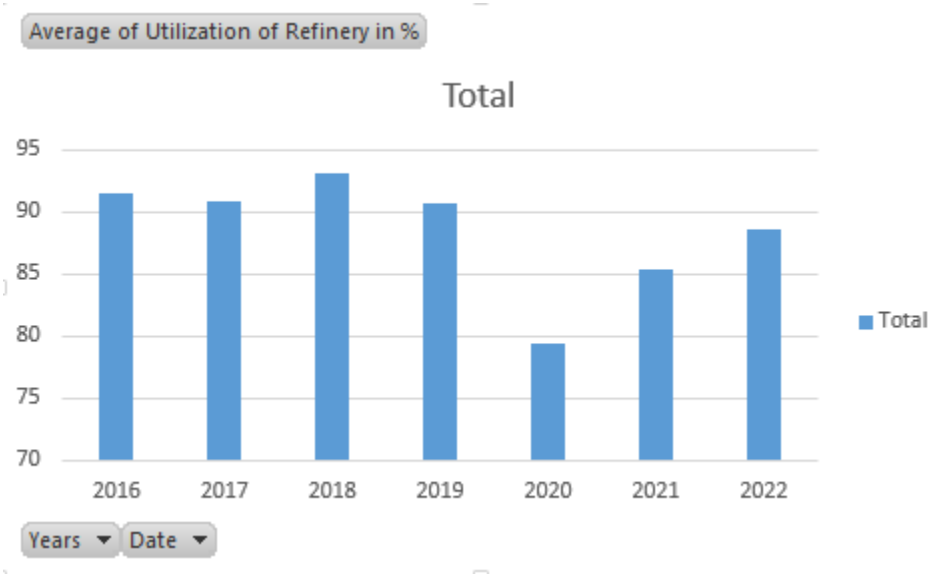


Figure 5 Refinery utilization

The refinery utilization refers to the value of maximum capacity percentage used in the processing of crude oil. The bar graph is showing that the maximum utilization of the refinery is seen in the year 2018 and the lowest utilization of the refineries are in the 2020.

Data preparation

The data collected is in time series format; therefore, time series forecasting models are employed, including auto-regressive integrated moving average (ARIMA) or multivariate time series model (VARMAX). In this data set, VARMAX is used for the multivariable analysis because it addresses the drawbacks of the ARIMA model by including the exogenous variables that influence the target variable which is the crude oil price in this case.

Prior to deploying the model, the abnormalities or outliers are found and eliminated. The rows that were containing missing values replaced with the constant value or mean values using linear interpolation technique.

```
oil_prices = oil_prices.resample('D').mean().pad()
oil_prices
```

Since this is a time series data set spanning the years 2016 to 2022 the anomalies are discovered using the approaches like interquartile range and Pycarat module, which incorporates the isolation forest method for identifying the anomalies (Su, et al., 2019). The outliers are removed by removing the rows that are generating abnormalities. Along with this Dickey-fuller test is used to determine if the data is stationary or non-stationary. Therefore, by applying augmented Dickey, the periodic tendencies or unpredictability in data is determined. For splitting the series into testing and training data, VAR model will be fitted on the data_training and then used for forecasting the upcoming 8 observations. Then the forecast will be compared with the actual values of the test data. For making this comparison, multiple forecast accuracy metrics is used.

```
nobs = 8
data_training, test_d = data_test_088[0:-nobs], data_test_088[-nobs:]
def adfuller_test(series, signif=0.05, name='', verbose=False):
    r = adfuller(series, autolag='AIC')
    output = {'test_statistic':round(r[0], 4), 'pvalue':round(r[1], 4), 'n_lags':round(r[2], 4), 'n_obs':r[3]}
    p_value = output['pvalue']
    def adjust(val, length= 6): return str(val).ljust(length)
    print(f'    Augmented Dickey-Fuller Test on "{name}"', "\n   ", '-'*47)
    print(f' Null Hypothesis: Data has unit root. Non-Stationary.')
    print(f' Significance Level    = {signif}')
    print(f' Test Statistic        = {output["test_statistic"]}')
    print(f' No. Lags Chosen       = {output["n_lags"]}')
    for key,val in r[4].items():
        print(f' Critical value {adjust(key)} = {round(val, 3)}')
    if p_value <= signif:
        print(f" => P-Value = {p_value}. Rejecting Null Hypothesis.")
        print(f" => Series is Stationary.")
    else:
        print(f" => P-Value = {p_value}. Weak evidence to reject the Null Hypothesis.")
        print(f" => Series is Non-Stationary.")
```

Chapter 5: Implementation of the learning models

The relationship between the dependent and independent variables is determined in this research work. The goal is to know what factors have a significant influence on crude oil prices in order to add those variables as exogenous variables in the model forecasts (Li, et al., 2019). In this scenario, the Granger casualty test is used to determine the relationship between the variables. In this, another hypothesis is established regarding the impact of different variables. In case the timeline or time series of independent and dependent variables are linked, the null hypothesis will be rejected as the P value internal hypothesis exceeds a specific threshold value.

Models

The forecast hypothesis is demonstrated in order to answer the cutting-edge question. The forecast will be focused on the news articles' prediction for the upcoming 2 or one years. The hypothesis will focus on whether the articles of the newspaper can effectively predict the prices and trends of crude oil. The key takeaways and sentiments from the news articles in relation to

the time series data set can be used to analyze the uncertainty and reaffirm the accuracy of news forecasts.

Time series modeling techniques

The literature review offered a number of techniques for creating time-series models. The methods include autoregressive neural networks, Holt winters, GARCH (generalized autoregressive conditional heteroscedastic), support vector regression, and ARIMA. A variety of hybrid models are also suggested including wavelets, ARIMA, genetic algorithms, and neural networks combines with the support vector regression (Sun, et al., 2022). The ARIMA, LSTM, and VAR approaches are discussed and compared. Then the VAR model is prepared for the prediction of the oil price.

ARIMA: ARIMA model is used in time-series based forecasting. This is the statistical method that is used for the analysis and development of the forecasting model that offers the best representation of the time series by using correlation modeling in the data. The ARIMA model seeks to capture the autocorrelations in the data. This is mainly used in the process of demand forecasting (Su, et al., 2019). This only needs the prior or historical data of the time series for the generalization of the forecast. This is mainly suitable for short-term forecasts. This is used for modeling the non-stationary time series.

VAR: when multiple time series impact each other the vector Autoregression which is the forecasting technique may be employed. That is the time series in the world have a bidirectional link. It is classified as an autoregressive model since every variable is modeled as a function of previous values and implies that the predictors are nothing more than the series' lags. Here, the term bi-directional refers that the variable's impact on one another (Kocaarslan, et al., 2020). Each variable in where the model is described as a linear mixture of its own past values of the other variables in the system

```
model = VAR(df_differenced)
for i in [1,2,3,4,5,6,7,8,9]:
    result = model.fit(i)
    print('Lag Order =', i)
    print('AIC : ', result.aic)
    print('BIC : ', result.bic)
    print('FPE : ', result.fpe)
    print('HQIC: ', result.hqic, '\n')
x = model.select_order(maxlags=12)
model_fitted = model.fit(2)
out = durbin_watson(model_fitted.resid)
for col, val in zip(data_test_088.columns, out):
    print((col), ':', round(val, 2))
lag_order = model_fitted.k_ar
forecast_input = df_differenced.values[-lag_order:]
fc = model_fitted.forecast(y=forecast_input, steps=nobs)
df_forecast = pd.DataFrame(fc, index=data_test_088.index[-nobs:], columns=data_test_088.columns + '_2d')
```

LSTM: This stands for Long short-term memory neural networks. It is a type of recurrent neural network. The key goal of this is to model the long-term dependencies and establish the best time lag for the problems that are related to time series. This is a sort of deep neural network architecture with a complex temporal structure. This is commonly utilized in the time series modeling process. A typical neural network is predicated on the assumption that all the units of input are not dependent on each other. As consequence, the sequential data cannot be used by the traditional neural network (Kocaarslan, and Soytas, 2019). The RNN model on the other hand includes a hidden state that is formed by sequential information from a time series along with the output reliant on the hidden state. The design of the LSTM neural network contains multiple hidden layers of delays which is the amount of the previous data points that are responsible for testing and training. By trial and error, the number of hidden layers and delays are adjusted to 5 and 4 respectively. The backpropagation technique is used to train the MLP and LSTM neural networks. The batch size the learning rate and the number of epochs are 32, 0.05, and 50 respectively. Along with time the learning rate decreases and it governs the face of convergence. The training may be converged by increasing the number of epochs and the learning rate to 1000 and 0.05 respectively.

```
[ ] model=Sequential()
    model.add(LSTM(units=128,activation='relu',input_shape=(X_train.shape[1], X_train.shape[2])))
    #model.add(Dropout(rate=0.2))
    model.add(Dense(units=1))
    model.compile(loss='mean_squared_error', optimizer='adam')

    WARNING:tensorflow:Layer lstm_5 will not use cuDNN kernels since it doesn't meet the criteria. It will use
```

```
[ ] history = model.fit(
        X_train, y_train,
        epochs=50,
        batch_size=32,
        validation_split=0.1,
        shuffle=False
    )
    .
    33/33 [==============================] - 1s 35ms/step - loss: 0.4661 - val_loss: 0.4053
    Epoch 11/50
    33/33 [==============================] - 1s 34ms/step - loss: 0.4661 - val_loss: 0.4053
```

Model selection criteria

The accuracy of the model may differ for different datasets. It is necessary to compare different factors and models determine the best suitable for the results with less mistakes. There are several approaches to calculating, measuring, and minimizing mistakes. Since no one metric can offer a precise indicator of prediction success and the use of several metrics makes it difficult to compare forecasting models and makes them uncontrollable, selecting metrics to remedy forecasting mistakes is sometimes a significant difficulty (Lu, et al., 2020). The following factors were determined to be appropriate from the extensive list of suitable model selection criteria that are MAPE ($Mean\ absolute\ percentage\ error$), MAE ($Mean\ Absolute\ error$), MSE ($Mean\ SquareError$), and RMSE ($Root\ Mean\ Square\ error$). The framework's over- or underestimation of the actual prediction value may likewise be inferred only inferentially from this assessment criterion. The second assessment criterion is RMSE. It keeps the estimated variable's units intact. This strategy minimizes significant mistakes and is more sensitive. However, using this criterion restricts the capacity to compare various time series. On the other hand, MAE, the third and final criterion, establishes the size of the error given a specific set of forecasts. How closely forecasts match actual results is determined by MAE. The projections' direction is not taken into account by this measure (Li, et al., 2019). Additionally, the accuracy of continuous variables is determined by these criteria. The fourth criterion, MAPE, allows for the comparison of various time-series data without specifying the relationship or % error. In situations where the observed variables are very big, this metric is important.

Chapter 6: Evaluation and results

In this research, an integrated model is built for forecasting the price of crude oil using a machine learning technique that is based on core factor selection. This research contributes to the selection of variable and machine-learning approaches for estimating the prices of crude oil. Furthermore, for making the oil price prediction LSTM approach is used. The model is combined with the ARIMA and VAR approaches (Norouzi, et al., 2020). Lastly, various outcomes are compared and contrasted using $mean\ absolute\ percentage\ error$ that is MAPE and $root\ mean\ squared\ error$ that is RMSE.

Evaluation criteria

In order to compare the ability and performance of forecasting the proposed approach with other benchmark models two key evaluation criteria are used in this research approach that is RMSE which is the $root\ mean\ squared\ error$ and MAPE which is the $mean\ absolute\ percentageerror$. These are used to measure performance. The null hypothesis states that the model under consideration has no ability to anticipate the crude oil prices.

Null hypothesis (H0): The model under the consideration has no ability to anticipate the crude oil prices as variables are independent to each other.

Alternative hypothesis (H1): The model under the consideration can effectively anticipate the crude oil prices and variables are dependent on each other.

Granger's causality test

Vector autoregression is based on the idea that every time series in the system affects the others. This allows us to anticipate the series based on its historical values along with the values of other time series of the system. The Grangers causality test is used to assess the relationship before developing the model (Hoang, et al., 2021). This inspects the null hypothesis that the regression equations coefficient of the past values is equal to zero. To put it another way, the previous values of the time series do not influence the other series. Therefore, if the P value produced from the test is less than the significant level of 0.05 then $the\ null\ hypothesis$ may be confidently rejected.The following output shows the granger's casualty test for all the possible

combinations of time series in the selected data frame and keeps the P values for each time series combination in the output matrix.

	Utilization of Refinery in %_x	Production_x	Demand_x	Contract 1 FP_x	Global Sale (in millions)_x	Last Day's Price_x	Price_x
Utilization of Refinery in %_y	1.0000	0.0000	0.0000	0.0000	0.0000	0.0076	0.0000
Production_y	0.0003	1.0000	0.3297	0.0000	0.0000	0.0223	0.0000
Demand_y	0.0010	0.0000	1.0000	0.0000	0.0000	0.0611	0.0000
Contract 1 FP_y	0.8569	0.0857	0.1843	1.0000	0.3458	0.0270	0.0839
Global Sale (in millions)_y	0.1989	0.0003	0.0000	0.0000	1.0000	0.0014	0.0000
Last Day's Price_y	0.9145	0.0746	0.1680	0.0000	0.2445	1.0000	0.0000
Price_y	0.8893	0.0685	0.1559	0.2805	0.2368	0.0339	1.0000

Figure 6 Granger's Causality test

Here the row signifies the response while the column represents the predictor series. When the P value is less in comparison to the significance level which is 0.05 then the corresponding column causes the row. For the production_x causing the Global sale, the p-value is 0.0003 which is less than 0.05. This allows us to reject the null hypothesis and conclude that the production_x causes Global sales. For the Last Day's Price_x causing the Global sale_y, the p-value is 0.0014 which is less than 0.05. This allows us to reject the null hypothesis and conclude that the Last Day's Price_x causes Global sales. For the Utilization of Refinery in %_x causing the Production_y, the p-value is 0.0003 which is less than 0.05. This allows us to reject the null hypothesis and conclude that the Utilization of Refinery in %_x causes Production_y. By considering the P values obtained from the above table it can be identified that most of the variables of time series in the system are causing each other indiscriminately. This makes the multi-time series system or strong option for forecasting with VAR models.

Cointegration test

The cointegration test is used to determine whether or not there is a statistically significant relationship between the time series. If there is a linear combination of multiple time series

considering the value of the order of integration is less than that of Individual series then the collection of time series can be referred to as cointegrated. The cointegration occurs when multiple time series have a statistically meaningful relationship (Su, et al., 2021). The results show that the utilization of the refinery in % and the production has test stat value 333.58 and 88.67 respectively which shows true significance.

```
Name   ::  Test Stat > C(95%)    =>   Signif
 ----------------------------------------
Utilization of Refinery in % ::  761.17    > 111.7797  =>   True
Production ::  442.38    > 83.9383   =>   True
Demand ::  189.48    > 60.0627   =>   True
Contract 1 FP ::  61.45     > 40.1749   =>   True
Global Sale (in millions) ::  20.53     > 24.2761   =>   False
Last Day's Price ::  7.18      > 12.3212   =>   False
Price  ::  0.03      > 4.1296    =>   False
```

Figure 7 Cointegration test

Checking for stationery

It is necessary that the time series must be stationary to prepare the VAR model. The stationary time series refers to the time series in which the value of the variance and the mean remain the same with time. There are multiple methods to test this such as the $ADF\ test, KPSS\ test,$ and $Philip\ perron\ test$. In this research, the ADF test is used. If the time series is discovered to be non-stationary then it is a must to make it stationary (Erickson, et. al., 2018). This can be done by differencing the series once and then repeating the test to make it stationary. Because the process of differencing decreases the length of the time series by 1 therefore, it is required that all the time series have the same length. For this, the $adfuller\_test()$ function is used.

```
   Augmented Dickey-Fuller Test on "Utilization of Refinery in %"
   ----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -3.4401
No. Lags Chosen       = 15
Critical value 1%     = -3.435
Critical value 5%     = -2.863
Critical value 10%    = -2.568
=> P-Value = 0.0097. Rejecting Null Hypothesis.
=> Series is Stationary.


   Augmented Dickey-Fuller Test on "Production"
   ----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -1.1015
No. Lags Chosen       = 24
Critical value 1%     = -3.435
Critical value 5%     = -2.863
Critical value 10%    = -2.568
=> P-Value = 0.7145. Weak evidence to reject the Null Hypothesis.
=> Series is Non-Stationary.


   Augmented Dickey-Fuller Test on "Demand"
   ----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -1.2285
No. Lags Chosen       = 24
Critical value 1%     = -3.435
```

Figure 8 ADF test

The results of the ADF test showed that the time series is not stationary. Therefore, all of these are different once again. After this, it has been identified that all the timeseries are stationary.

```
   Augmented Dickey-Fuller Test on "Utilization of Refinery in %"
   ----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -11.5427
No. Lags Chosen       = 20
Critical value 1%     = -3.435
Critical value 5%     = -2.863
Critical value 10%    = -2.568
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.


   Augmented Dickey-Fuller Test on "Production"
   ----------------------------------------------
Null Hypothesis: Data has unit root. Non-Stationary.
Significance Level    = 0.05
Test Statistic        = -11.7132
No. Lags Chosen       = 23
Critical value 1%     = -3.435
Critical value 5%     = -2.863
Critical value 10%    = -2.568
=> P-Value = 0.0. Rejecting Null Hypothesis.
=> Series is Stationary.


   Augmented Dickey-Fuller Test on "Demand"
   ----------------------------------------------
```

Figure 9 Stationary time series

Selecting the order of the VAR model

For choosing the proper order of the VAR model, increase orders of the VAR model repeatedly and choose the order that produces the model with the lowest AIC value. However, the AIC is the most commonly used for the analysis, the BIC, HQIC, and FPE have also been considered the key comparison estimates.

```
Lag Order = 1
AIC :  28.743283192799687
BIC :  28.932370942382914
FPE :  3041231919984.9434
HQIC:  28.813517614926564

Lag Order = 2
AIC :  27.95336861759587
BIC :  28.308089608789267
FPE :  1380369780418.216
HQIC:  28.085129718121305

Lag Order = 3
AIC :  27.538344304087683
BIC :  28.05886819687979
FPE :  911499325171.4869
HQIC:  27.73169899173442

Lag Order = 4
AIC :  27.21058915085691
BIC :  27.897085889072805
FPE :  656781730120.9946
HQIC:  27.46560444944623

Lag Order = 5
AIC :  26.894539207810464
BIC :  27.747179019765507
FPE :  478821492512.964
HQIC:  27.21128225653183
```

Figure 10 Lag order

From the above output, it is identified that the value of the AIC drops to the lag order 9. Alternatively, $the\ model.select\_order(maxlags)$ method is used as it gives the value of the selected order that has the lowest $BIC, AIC, HQIC, and\ FPE$.

VAR Order Selection (* highlights the minimums)

	AIC	BIC	FPE	HQIC
0	30.92	30.94	2.671e+13	30.92
1	28.79	28.98	3.190e+12	28.86
2	28.00	28.35	1.441e+12	28.13
3	27.58	28.10	9.477e+11	27.77
4	27.25	27.93	6.798e+11	27.50
5	26.93	27.78	4.937e+11	27.24
6	26.74	27.76	4.089e+11	27.12
7	26.62	27.81	3.656e+11	27.07
8	26.48	27.84	3.177e+11	26.99
9	26.29	27.81	2.611e+11	26.85
10	26.05	27.73	2.049e+11	26.67
11	25.88	27.73*	1.734e+11	26.57
12	25.73*	27.75	1.499e+11*	26.48*

Figure 11 Value order selection

As the BIC and HQIC are lowest at lag 2, the selected order is chosen 2 for the model.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Thu, 17, Nov, 2022
Time:                     10:06:37
--------------------------------------------------------------------
No. of Equations:         7.00000    BIC:                    28.3081
Nobs:                     1590.00    HQIC:                   28.0851
Log likelihood:          -37910.7    FPE:                 1.38037e+12
AIC:                      27.9534    Det(Omega_mle):      1.29256e+12
--------------------------------------------------------------------
```

Figure 12 Summary of regression results

Serial Correlation of the errors

This approach is sued to determine the presence of the leftover pattern in the errors. This refers that if there is a connection in the errors, the model still has to be able to comprehend some pattern in the time series. The conventional course of action in such a situation is to either add additional predictors in the system, raise the model's order or seek an alternative method to describe the time series. This is used to ensure that the model effectively explains the patterns and variances in the time series. Following is the formula of $Durbin\ Watson's$ statistic:

$$DW = \frac{\Sigma_{t=2}^{T}((e_t - e_{t-1})^2)}{\Sigma_{t=1}^{T} e_t^2}$$

The statistic's value can range from 0 to 4. There is less of a meaningful serial connection the closer it gets to the value of 2. When the value is near 0, then the positive correlation can be determined, and when the value is above 2.0 then the negative correlation is identified.

```
Utilization of Refinery in % : 2.11
Production : 2.19
Demand : 2.17
Contract 1 FP : 2.14
Global Sale (in millions) : 2.05
Last Day's Price : 2.16
Price : 2.12
```

Figure 13 Serial correlation

Forecasting performance evaluation

Three models are chosen to test the prediction potential of the proposed machine learning integrated approach including $ARIMA, VAR, and\ LSTM$. According to the results, the $variable\ selection\ machine\ learning\ approach$ in the testing and training set, what's better than the single variable precision model as well as the $core\ factors\ extracted\ model$ (Abdollahi, and Ebrahimi,2020).

The VAR model anticipates up to the observation's lag order from the historical data for making the forecast. This is due to the fact that the terms in the VAR model are primarily the lags of the different time series in the dataset, therefore, it is necessary to supply it with as many of the prior values as the lag order suggested by the model requires.

```
array([[ 0.        ,  0.63671937,  0.73531621,  0.        ,  0.        ,
        -1.41999817,  0.72999573],
       [ 0.        ,  0.63671937,  0.73531621,  0.        ,  0.        ,
         0.72999573, -0.37999725]])
```

Figure 14 Forecast input

Date	Utilization of Refinery in %_forecast	Production_forecast	Demand_forecast	Contract 1 FP_forecast	Global Sale (in millions)_forecast	Last Day's Price_forecast	Price_forecast
11/2/2022	74.555813	4543.114187	4171.943176	77.215038	4.661227	47.934157	39.331300
11/3/2022	67.268123	4582.296213	4102.086902	82.019293	3.999192	48.029083	37.110277
11/4/2022	59.060120	4673.535069	3954.471553	89.509467	3.550907	47.607814	34.169520
11/5/2022	52.050859	4769.778411	3874.435694	96.361472	3.002039	44.076415	30.096650
11/6/2022	44.717233	4842.979665	3773.336984	102.582985	2.483761	43.873039	27.517851
11/7/2022	37.347643	4931.090673	3661.084297	109.277254	1.930477	42.257558	24.254731
11/8/2022	29.953746	5016.858002	3564.206979	115.973778	1.433514	40.208931	20.900574
11/9/2022	22.608272	5098.335156	3461.200053	122.525339	0.904052	39.182299	17.793985

Figure 15 Forecast

Forecasts are produced but there are limited by the size of the model's training set. Therefore, it is required to de-difference this equal number of times to the original input data's differences to return this to the original scale.

Inverting the transformation

This is done to bring the forecast back to the original scale. The following graph shows the plot of the forecast against the actual values of the variables.

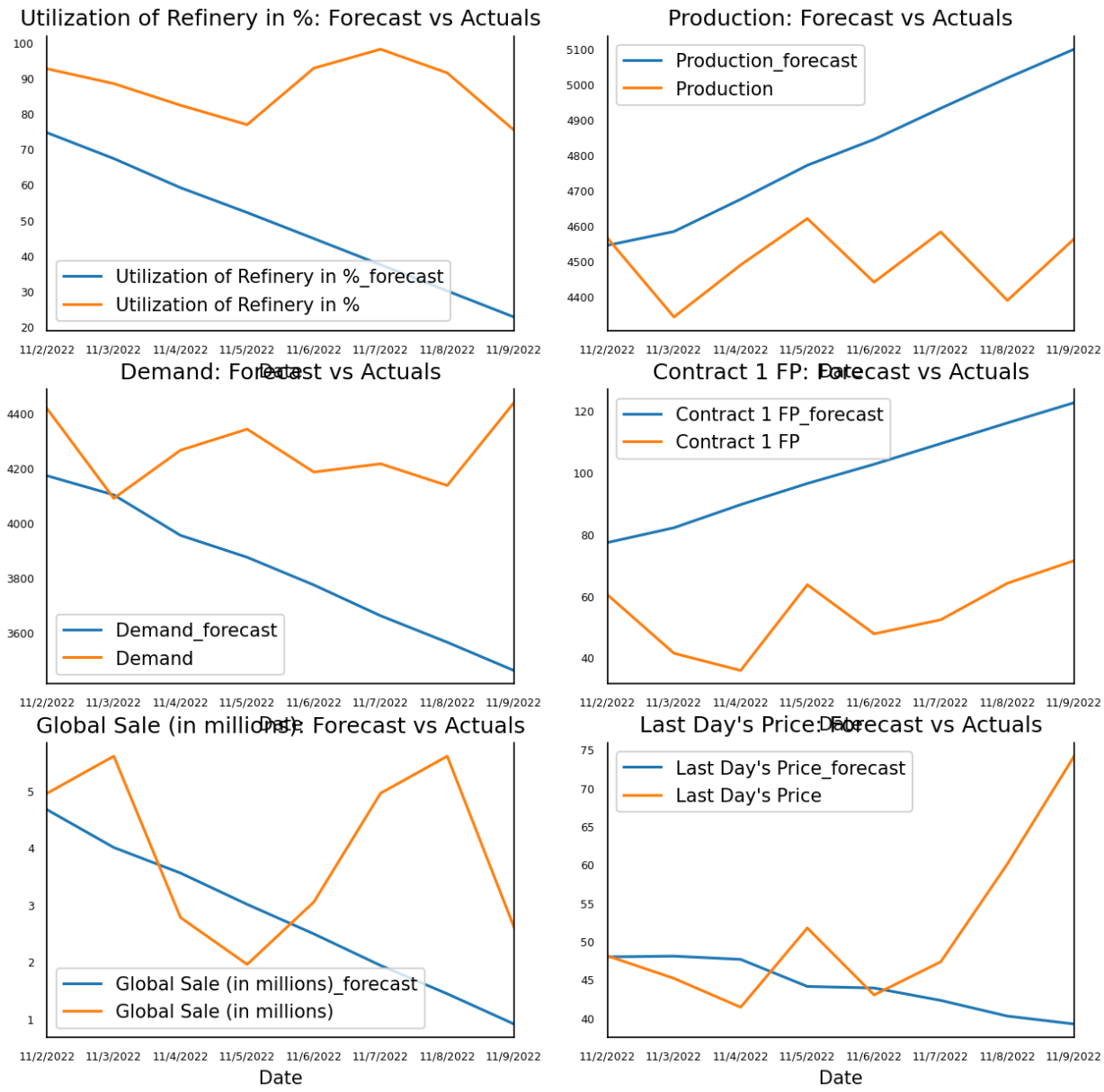


Figure 16 Forecast plot

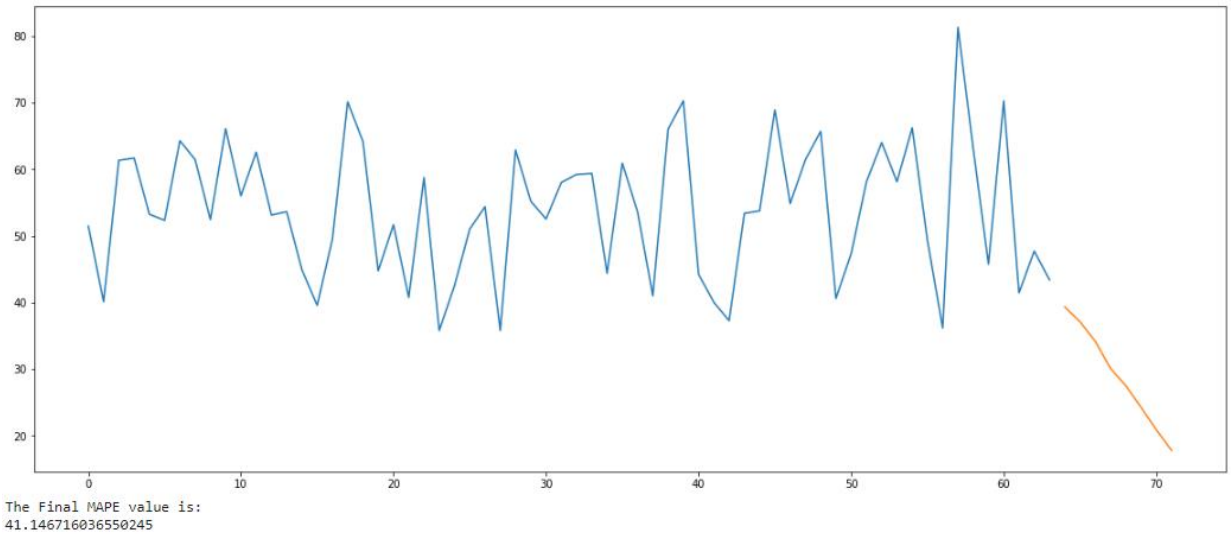


Figure 17 Forecast evaluation of Price using MAPE

```
Forecast Accuracy of: Price
mape :  0.4115
me :  -23.7994
mae :  23.7994
mpe :  -0.4115
rmse :  27.8363
corr :  0.0047
minmax :  0.4115
```

Figure 18 Forecast accuracy of Price

The MAPE value for the accuracy of the price forecasting model is 0.4115, ME value is -23.799, MAE value is 23.799, MPE value is -0.41, RMSE value is27.83, Correlation is 0.0047 and Minmax is 0.4115. The final MAPE value of the price prediction model is 41.14. Thus, this allows rejecting the null hypothesis and adopting the alternative hypothesis as the model under consideration is able to anticipate the values of the crude oil.

The VAR model is prepared along with conducting the causality test, finding the optimal number of the VAR model, building the model, seeking serial autocorrelation, and so on. The results are compared with the accuracy metrics. The VAR predictions often beat the futures and random walk forecasts and have the least average forecast residuals and the best accuracy. This suggests using the random walk or futures with caution when making forecasts, especially for short-term projections (Mundia, et al., 2019).

Statistics test

The suggested technique outperforms all the other standard models with a confidence level of 99%. One probable explanation is that the proposed machine learning integration model considerably enhances the prediction performance of the model. When the LR model is used for testing the results are considered less than -7.015, which shows that the predictive performance of the integrated techniques is better in comparison to the machine learning model without the selection of variable (Gupta, and Nigam, 2020).It has been determined data variable selection machine learning technique which is based on LSTM, has the best direction prediction performance while the ARIMA forecast performance is the poorest.

Discussion

One of the most significant commodities in the world, crude oil is responsible for almost one-third of the world's energy use. It serves as the foundation of the majority of the items used by people on the daily basis such as plastic to fuels. Since the changes in the price of crude oil have a significant influence on the national economies throughout the world, price forecasting can be helpful to reduce the risks brought on by oil price volatility. For a variety of stakeholders, the price projection of crude oil is crucial. The structural models are those that take into account economic factors such as demand and supply. The price of crude oil is influenced by a variety of different factors, despite the fact that the structural models are thought to be the most rational ones for modeling the pricing of industrial goods (Yang, 2019). Other types of models referred to as time series models are non-structural and take crude oil price volatility into account over time. The time series data for crude oil prices is readily available and so makes it simpler to develop a time series model, it is hard to get trustworthy data for the formulation of the structural model. Therefore, in this work, a time series model is prepared for the analysis and prediction of crude oil prices.

The present price of crude oil is considered in the time series models to determine the impact of all the contributing variables and price forecasting may be done using the historical behavior of the prices of crude oil. The primary premise of such models is that the oil price history may predict future prices. Time series models have limits in their capacity to forecast when trend inversions are shown in the data or when the recurring pattern caught in the model is not followed in the future pricing, despite the fact that these models may capture the pattern and

trends or the other repetitive patterns in the data (Cen, and Wang, 2019). Applying the three models to the time series of Normal *West Texas Intermediate* (WTI) crude oil prices allowed for comparison. The results of the comparison showed that given a prediction with a 95% confidence interval, the proposed VAR model outperformed. The model has the greatest results which led to drawing the conclusion that it beat other straightforward yet adaptable models in the oil market. In the overall process of advising policymakers, forecasting is essential. When the changes are assessed from two angles- current occurrences and what is expected to happen in the future is expected to happen in the future, make perfect decision-making results. Regardless of how hazy a prediction may appear, policymakers are required to take its veracity into account when making decisions (Yu, et al., 2017). To tighten the regulations and produce final results that deviate from the expected outcome, policymakers should ideally base their actions on reliable projections. The MAPE value for the accuracy of the price forecasting model is 0.4115, ME value is -23.799, MAE value is 23.799, MPE value is -0.41, RMSE value is27.83, Correlation is 0.0047 and Minmax is 0.4115. The final MAPE value of the price prediction model is 41.14.Therefore, the null hypothesis can be rejected as the VAR model can accurately anticipate the price of the crude oil. Additionally, for planning, devising, and implementing marketing plans, marketing strategists collect data from a variety of forecasting models and take the most accurate models into consideration. Consequently, the organizations whose primary functions rely on the oil sector find such projections particularly helpful for creating their marketing plans and policy plans to accommodate the future changes predicted by the models. Companies involved in oil production can utilize precise projections to use the data to make informed decisions about the pricing they assign to oil (Abdollahi, and Ebrahimi,2020). As a consequence, they may alter oil prices in a way that avoids harming the financial capacity or otherwise influencing the goals of the organizations by having the correct information on the flow of public money as well as on the specific patterns in oil consumption and demand.

Chapter 7: Conclusion

In this research work, a machine learning and variable selection framework is developed for forecasting the crude oil price that integrates the variable selection and LSTM. In the analysis, the data from the year 2016 to 2022 is considered. The performance of the model for forecasting is compared to other major approaches such as $ARMA, LSTM, and\ VAR$. The dickey-fuller test is used to determine if the data is stationary or non-stationary. In both the testing and training sets the variable choice machine learning integration approach suggested in this research work outperformed the model without the core factor extraction and Univariate model. The predictive ability of the different models is evaluated. The results demonstrated that the proposed model's prediction accuracy is much higher than that of the univariate model. The number of the key variables chosen by the BMA is neither the greatest nor the smallest, which shows that the number of core variables also influences the prediction outcomes. It has been identified that the forecasting performance changes over time. It has been identified the variable selection-based machine learning approach using this resource study enhances the oil price predicting ability substantially. The performance of the VAR model, despite the overall strength, can be unstable throughout the whole sample with slight changes in the specifications or lag length occasionally leading to wildly divergent projections. It is concluded that there is still difficulty in forecasting the oil prices on the large sample of the data with structural breakdowns and finding benefits in the combination prediction for longer time series. In a future study, more independent variables may be included in the resource using Internet search data to assess the efficacy of the presented approach. This approach also allows for the quantification of investor sentiment. Furthermore, several variable selection approaches may be incorporated into this.

References

Peng, J., Li, Z. and Drakeford, B.M., 2020. Dynamic characteristics of crude oil price fluctuation—from the perspective of crude oil price influence mechanism. *Energies*, *13*(17), p.4465.

Khan, K., Su, C.W., Umar, M. and Yue, X.G., 2021. Do crude oil price bubbles occur?. *Resources Policy*, *71*, p.101936.

Orojo, O., Tepper, J., McGinnity, T.M. and Mahmud, M., 2019, December. A multi-recurrent network for crude oil price prediction. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 2940-2945). IEEE.

Wen, F., Min, F., Zhang, Y.J. and Yang, C., 2019. Crude oil price shocks, monetary policy, and China's economy. *International Journal of Finance & Economics*, *24*(2), pp.812-827.

Abdollahi, H. and Ebrahimi, S.B., 2020. A new hybrid model for forecasting Brent crude oil price. *Energy*, *200*, p.117520.

Lu, Q., Li, Y., Chai, J. and Wang, S., 2020. Crude oil price analysis and forecasting: A perspective of "new triangle". *Energy Economics*, *87*, p.104721.

Li, X., Shang, W. and Wang, S., 2019. Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, *35*(4), pp.1548-1560.

Gupta, N. and Nigam, S., 2020. Crude oil price prediction using artificial neural network. *Procedia Computer Science*, *170*, pp.642-647.

Su, C.W., Huang, S.W., Qin, M. and Umar, M., 2021. Does crude oil price stimulate economic policy uncertainty in BRICS?. *Pacific-Basin Finance Journal*, *66*, p.101519.

Sun, C., Zhan, Y., Peng, Y. and Cai, W., 2022. Crude oil price and exchange rate: Evidence from the period before and after the launch of China's crude oil futures. *Energy Economics*, *105*, p.105707.

Cen, Z. and Wang, J., 2019. Crude oil price prediction model with long short term memory deep learning based on prior knowledge data transfer. *Energy*, *169*, pp.160-171.

Aloui, D., Goutte, S., Guesmi, K. and Hchaichi, R., 2020. COVID 19's impact on crude oil and natural gas S&P GS Indexes.

Beckmann, J., Czudaj, R.L. and Arora, V., 2020. The relationship between oil prices and exchange rates: Revisiting theory and evidence. *Energy Economics*, *88*, p.104772.

Corbet, S., Goodell, J.W. and Günay, S., 2020. Co-movements and spillovers of oil and renewable firms under extreme conditions: New evidence from negative WTI prices during COVID-19. *Energy economics*, *92*, p.104978.

Erickson, P., Lazarus, M. and Piggot, G., 2018. Limiting fossil fuel production as the next big step in climate policy. *Nature Climate Change*, *8*(12), pp.1037-1043.

Hoang, A.T., Nižetić, S., Olcer, A.I., Ong, H.C., Chen, W.H., Chong, C.T., Thomas, S., Bandh, S.A. and Nguyen, X.P., 2021. Impacts of COVID-19 pandemic on the global energy system and the shift progress to renewable energy: Opportunities, challenges, and policy implications. *Energy Policy*, *154*, p.112322.

Johra, H., Heiselberg, P. and Le Dréau, J., 2019. Influence of envelope, structural thermal mass and indoor content on the building heating energy flexibility. *Energy and Buildings*, *183*, pp.325-339.

Kocaarslan, B. and Soytas, U., 2019. Asymmetric pass-through between oil prices and the stock prices of clean energy firms: New evidence from a nonlinear analysis. *Energy Reports*, *5*, pp.117-125.

Kocaarslan, B., Soytas, M.A. and Soytas, U., 2020. The asymmetric impact of oil prices, interest rates and oil price uncertainty on unemployment in the US. *Energy Economics*, *86*, p.104625.

Mundia, C.W., Secchi, S., Akamani, K. and Wang, G., 2019. A regional comparison of factors affecting global sorghum production: The case of North America, Asia and Africa's Sahel. *Sustainability*, *11*(7), p.2135.

Norouzi, N., de Rubens, G.Z., Choupanpiesheh, S. and Enevoldsen, P., 2020. When pandemics impact economies and climate change: Exploring the impacts of COVID-19 on oil and electricity demand in China. *Energy Research & Social Science*, *68*, p.101654.

Obringer, R., Mukherjee, S. and Nateghi, R., 2020. Evaluating the climate sensitivity of coupled electricity-natural gas demand using a multivariate framework. *Applied Energy*, *262*, p.114419.

Solarin, S.A., 2020. An environmental impact assessment of fossil fuel subsidies in emerging and developing economies. *Environmental Impact Assessment Review*, *85*, p.106443.

Su, C.W., Huang, S.W., Qin, M. and Umar, M., 2021. Does crude oil price stimulate economic policy uncertainty in BRICS?. *Pacific-Basin Finance Journal*, *66*, p.101519.

Su, C.W., Khan, K., Tao, R. and Nicoleta-Claudia, M., 2019. Does geopolitical risk strengthen or depress oil prices and financial liquidity? Evidence from Saudi Arabia. *Energy*, *187*, p.116003.

Umar, M., Su, C.W., Rizvi, S.K.A. and Lobonţ, O.R., 2021. Driven by fundamentals or exploded by emotions: Detecting bubbles in oil prices. *Energy*, *231*, p.120873.

Watts, N., Amann, M., Arnell, N., Ayeb-Karlsson, S., Beagley, J., Belesova, K., Boykoff, M., Byass, P., Cai, W., Campbell-Lendrum, D. and Capstick, S., 2021. The 2020 report of the Lancet Countdown on health and climate change: responding to converging crises. *The Lancet*, *397*(10269), pp.129-170.

. Connectedness of economic policy uncertainty and oil price shocks in a time domain perspective. *Energy Economics*, *80*, pp.219-233.

Zhao, L., Cheng, L., Wan, Y., Zhang, H. and Zhang, Z., 2015. A VAR-SVM model for crude oil price forecasting. *International Journal of Global Energy Issues*, *38*(1-3), pp.126-144.

Zhao, Y., Li, J. and Yu, L., 2017. A deep learning ensemble approach for crude oil price forecasting. *Energy Economics*, *66*, pp.9-16.

Zhang, J.L., Zhang, Y.J. and Zhang, L., 2015. A novel hybrid method for crude oil price forecasting. *Energy Economics*, *49*, pp.649-659.

Miao, H., Ramchander, S., Wang, T. and Yang, D., 2017. Influential factors in crude oil price forecasting. *Energy Economics*, *68*, pp.77-88.

Yu, L., Zhang, X. and Wang, S., 2017. Assessing potentiality of support vector machine method in crude oil price forecasting. *EURASIA Journal of Mathematics, Science and Technology Education*, *13*(12), pp.7893-7904.