Crude oil price prediction

1.AIM

To design a machine learning model that is able to produce crude oil price prediction

Oil demand is inelastic, therefore the rise in price is good news for producers because they will see an increase in their revenue. Oil importers, however, will experience increased costs of purchasing oil. Because oil is the largest traded commodity, the effects are quite significant. A rising oil price can even shift economic/political power from oil importers to oil exporters. The crude oil price movements are subject to diverse influencing factors.

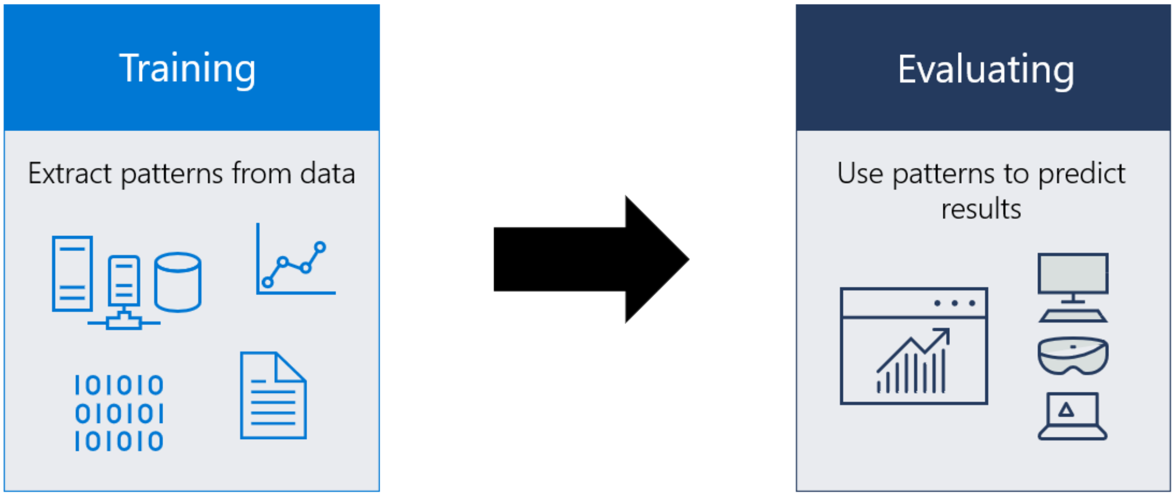
This Guided Project mainly focuses on applying Neural Networks to predict the Crude Oil Price.This decision helps us to buy crude oil at the proper time. Time series analysis is the best option for this kind of prediction because we are using the Previous history of crude oil prices to predict future crude oil. So we would be implementing RNN(Recurrent Neural Network) with LSTM(Long Short Term Memory) to achieve the task.
The volatility of the crude oil market and its chain effects to the world economy augmented the interest and fear of individuals, public and private sectors. Previous statistical and econometric techniques used for prediction, offer good results when dealing with linear data. Nevertheless, crude oil price series deal with high nonlinearity and irregular events. The continuous usage of statistical and econometric techniques for crude oil price prediction might demonstrate demotions to the prediction performance. Machine Learning and Computational Intelligence approach through combination of historical quantitative data with qualitative data from experts' view and news is a remedy proposed to predict this.

Crude oil is one of the most powerful resources in the world. The fluctuation of crude oil price plays an important role in the development of bulk commodities and the global economy. Under the comprehensive effects of market supply and demand game, US dollar exchange rate, speculative trading, geographical conflicts, natural disasters and other factors, the international crude oil price fluctuates sharply, which increases the difficulty of crude oil price prediction. Therefore, to build a scientific and reasonable model to accurately predict the trend of international crude oil price has become a hot and difficult issue in academic circles, investment circles and political circles.

However, due to the comprehensive effects of factors mentioned above, the fluctuation of crude oil price presents nonstationarity and nonlinearity making the prediction of crude oil price a challenging task. The research of crude oil price forecasting mainly includes two directions. The first direction is choosing effective models or improving the algorithm to better extract the features of price series and then predict. The second direction is to find the external indicators that affect the crude oil prices series, including financial policy, the price of related financial products, news sentiment and public opinions, to better predict the future trend of the original series.

In recent years, a novel “decomposition and ensemble” framework has been widely used in time series prediction, which can significantly improve the forecasting accuracy. In that framework, the original sequence is first decomposed into several components, and then each component is predicted by a single model.

Fig1.1 Training and Evaluation



Finally, the several prediction results are integrated to get the final prediction result. For the second direction, some researchers found that news articles and social media data were pretty beneficial in financial prediction. And other research methods proved that crude oil price had a significant relationship with different economic indicators. Then they used the Empirical Mode Decomposition (EMD) method to study the relationship between Gross Domestic Product (GDP) of the US and crude oil prices. found that political events and economic news, the same as oil supply and demand, played an important role in oil prices.

In the financial market, information sentiment contained in news articles and social media data is an important index reflecting the sentiment and viewpoint of investors and traders. The text contents of news include not only the report of facts, but also the intonation of language and emotion. Therefore, the news describing the fluctuation of crude oil price reflects the crude oil market situation through texts and influences the investor sentiment through the way of network communication. However, the consideration of these text data makes the analysis of the financial market even more complex . Inspired by this correlation, we quantify crude oil news as a sentiment index and introduce it into crude oil price prediction models.

2.RELATED WORK

In recent years, a large number of prediction models have been proposed. It can be divided into three categories: time series models, Artificial Intelligence (AI) models and hybrid prediction models. For the first categories established the Autoregressive Integrated Moving Average (ARIMA) model to quantitatively predict the international oil price. Then few groups proved that Vector AutoRegression (VAR) models could have good performance when forecasting short-term crude oil prices.

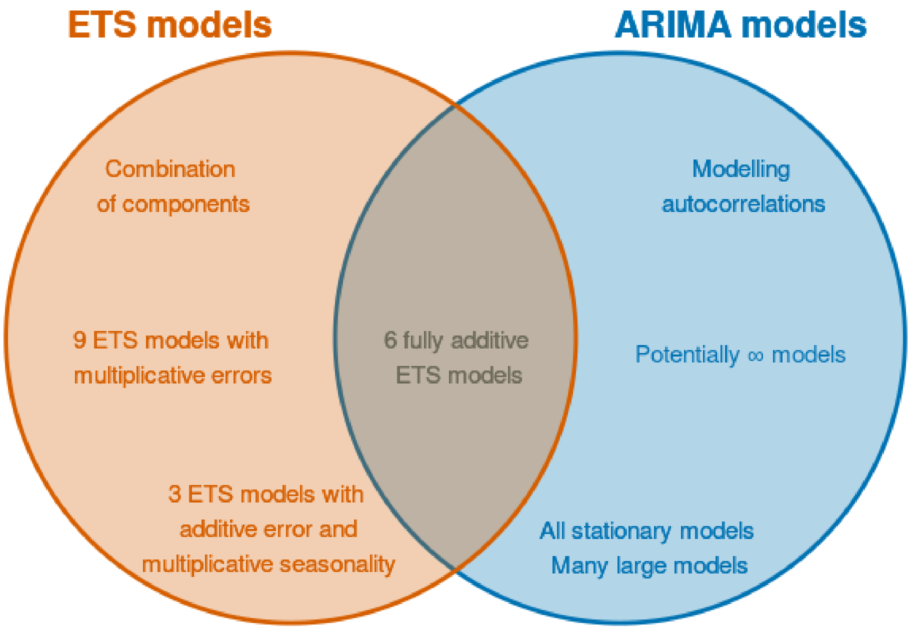
Although the time series model can better describe the linear characteristics of crude oil price series, it is hard to fit the nonlinear characteristics of crude oil price series. Therefore, many researchers introduced AI models into oil price prediction. Many used neural network models to forecast the monthly price of WTI crude oil from January 1992 to June 2008. utilized Least Squares Support Vector Regression (LSSVR) model to predict US WTI crude oil price.

Taking gold prices into consideration in the forecasting, few groups built a multiple wavelet Recurrent Neural Network (RNN) model for crude oil price forecasting. The experimental results showed the effectiveness of the model. Deep neural network model to the price prediction of WTI crude oil and achieved good results. Because of the multiple characteristics of the crude oil price series, the mixed model with different models became an effective choice.

Later few proposed network approaches had the ability to improve the prediction results for both spot oil prices and future oil prices. which integrated exponential smoothing model, ARIMA model and nonlinear autoregressive neural network.

Owing to nonstationarity and nonlinearity of the original price series, the family of EMD provides a new method for processing time series data. It starts from the characteristics of data itself and reveals the internal fluctuation characteristics of data by decomposing the fluctuation information of the original signal on different scales. Some researchers have demonstrated that it is an effective time series analysis tool and applied it to price forecasting.

Fig 2.1 ARIMA models

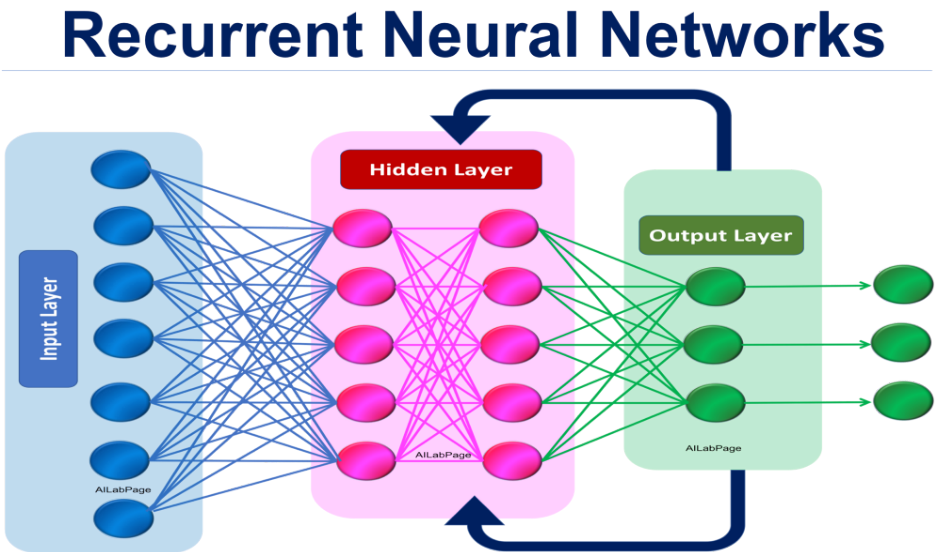

ETS models
ARIMA models
Combination of components
9 ETS models with multiplicative errors
3 ETS models with additive error and multiplicative seasonality
6 fully additive ETS models
Modelling autocorrelations
Potentially ∞ models
All stationary models
Many large models

For the consideration of text data, many studies have proved the correlation between investor sentiment and stock market volatility.Few researchers used the emotional tendency of financial reviews to predict future financial trends.

In recent years, a large number of prediction models have been proposed. It can be divided into three categories: time series models, Artificial Intelligence (AI) models and hybrid prediction models. For the first categories] established the Autoregressive Integrated Moving Average (ARIMA) model to quantitatively predict the international oil price. THen they proved that Vector AutoRegression (VAR) models could have good performance when forecasting short-term crude oil prices.

Although the time series model can better describe the linear characteristics of crude oil price series, it is hard to fit the nonlinear characteristics of crude oil price series. Therefore, many researchers introduced AI models into oil price prediction. Later some researchers used a neural network model to forecast the monthly price of WTI crude oil from January 1992 to June 2008. utilization of Least Squares Support Vector Regression (LSSVR) model to predict US WTI crude oil price. And taking gold prices into consideration in the forecasting built a multiple wavelet Recurrent Neural Network (RNN) model for crude oil price forecasting.

Fig 2.2 Recurrent Neural Network



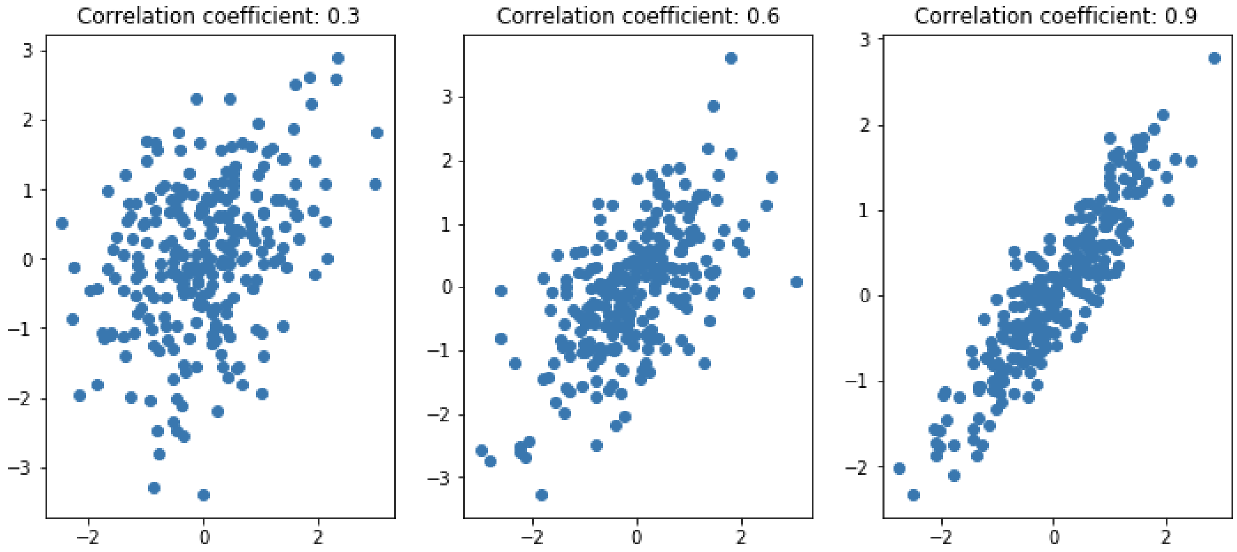
The experimental results showed the effectiveness of the model applied deep neural network model to the price prediction of WTI crude oil and achieved good results. Because of the multiple characteristics of the crude oil price series, the mixed model with different models became an effective choice.

To improve this a proposed network approach had the ability to improve the prediction results for both spot oil prices and future oil prices a hybrid forecasting model which integrated exponential smoothing model, ARIMA model and nonlinear autoregressive neural network.

Owing to nonstationarity and nonlinearity of the original price series, the family of EMD provides a new method for processing time series data. It starts from the characteristics of data itself and reveals the internal fluctuation characteristics of data by decomposing the fluctuation information of the original signal on different scales. Some researchers have demonstrated that it is an effective time series analysis tool and applied it to price forecasting.

For the consideration of text data, many studies have proved the correlation between investor sentiment and stock market volatility using the emotional tendency of financial reviews to predict future financial trends. proved that there was a high positive correlation between stock index and online sentiment analysis using linear regression tracking the public mood state from the content of huge amounts of micro-blog feeds by simple text processing techniques.

Fig 2.3 Correlation coefficient



However, in the field of crude oil market, there is little research on news sentiment analysis and crude oil price fluctuation. utilized the VAR model to study the price fluctuation of the global crude oil market, which showed that news sentiment has an important influence on the fluctuation of crude oil price. They also demonstrated news sentiment not only has an impact on the noise residual of oil, but also on the basic price trend through the regression analysis of news sentiment and oil time series decomposition components.

3.METHODS USED

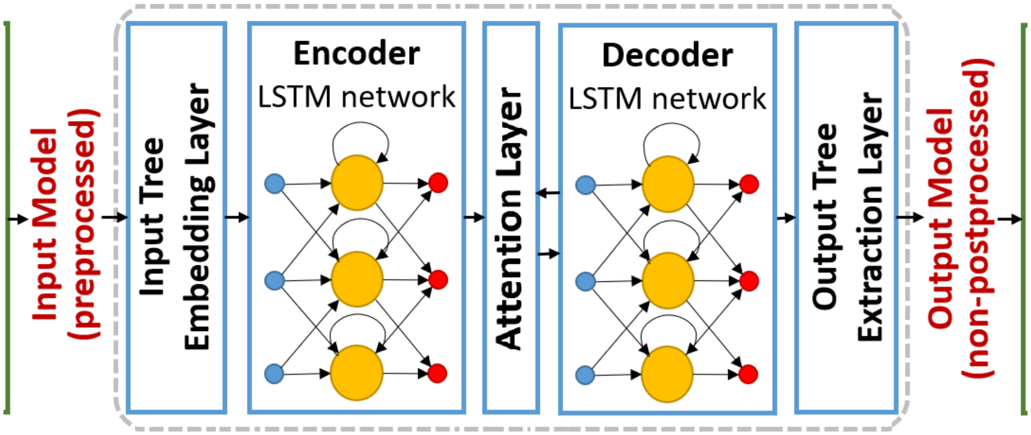
The method of using an LSTM model and other two comparable models, i.e., the ARIMA model and ANN model. Traditional time-series models have made great progress in the field of research on financial market prediction, especially the ARIMA model. As a nonlinear method in the field of artificial intelligence, ANNs can deal with nonlinear, discontinuous and high-frequency multidimensional data, and they have been widely used in financial forecasting. Therefore, in this section, we present the ARIMA and ANN models as the reference methods.

3.1. LSTM model

The LSTM model is a new type of deep machine learning neural network based on recurrent neural network (RNN) models. Hochreiter and Schmidhuber (1997) first came up with and improved this model, and then Graves (2012) extended it, with focus on the process of deep learning and development.

One of the key advantages of RNNs is that they can connect the previous information to the current task, that is, the previous information can be memorized and applied for the calculation of the current output, and the time-series data can be analyzed and predicted.

Fig 3.1.1 LSTM model



LSTM mainly solves the problems of RNNs, such as gradient disappearance and gradient explosion . LSTM constructs a long-term delay between input, output and gradient burst prevention. It has its own memory and can yield accurate predictions of crude oil prices. Besides that, LSTM models have excellent long-term and short-term memory ability, which will not lead to the loss of more historical state information on crude oil prices.

It can fully extract historical information on the crude oil price and consider the characteristics of the current data for price information. LSTM models have great advantages in terms of mining the long-term dependence of crude oil price sequence data.

Furthermore, LSTM models can automatically search for nonlinear features and complex patterns of crude oil prices, which shows excellent forecasting performance in crude oil price prediction.

As a very powerful prediction tool, LSTM has been widely used in prediction-related fields. Therefore, in order to forecast crude oil price more accurately.

The cell structure of LSTM consists of a cell state and three gates, which are the input, forget and output gates. The cell state records the state of neurons through memory storage, which is the core of the LSTM unit structure. However, whether the cell state is remembered depends on the control gate, which allows crude oil price information to be selectively transmitted

It has the advantages of adding information to the cell state or removing the information so as to defend and regulate the cell state. It contains a sigmoid layer and a pointwise multiplying computation. They use an LSTM model to conduct this forecasting analysis. In the construction of the LSTM model, a batch-normalization (BN) layer and dropout layer were added to optimize the structure of the neural network.

There are two problems, which may directly affect the training capability.

One is the problem of gradient disappearance, which makes model convergence difficult. The other is that the tests may fail because of overfitting. As for these problems, BN can effectively address the problem of gradient disappearance, and dropout technology can alleviate the problem of overfitting.

For this study, the LSTM model consists of three LSTM neural layers and two others closely connected ones. Each LSTM layer consists of 200 nodes. Before each LSTM neural layer, a BN layer was added, followed by a dropout layer; the inactivation probability was set as 0.2. Furthermore.

We used a mini-batch method to train the LSTM network and selected the mean squared error (MSE) as the loss function. Compared with other algorithms, the adaptive moment estimation (Adam) algorithm has the advantages of faster convergence speed and a better learning effect. Therefore, the Adam optimizer was selected for optimization training.

3.2.GLOBAL ECONOMIC DEVELOPMENT

Global economic development is a manifestation of supply and demand measuring the global demand shock directly by correcting the real gross domestic product (GDP) growth forecast. The results showed that the forecast was associated with unexpected growth in emerging economies during the 2022-2023 period.

These surprises were associated with a hump-shaped response of the real price of oil that reaches its peak after 12–16 months. The global real economic activity has always been considered to impact the changes in oil price significantly. When researched the relationship between global economic and oil prices with trend and cycle decomposition.

They found that economic shock has a lasting effect on oil prices, which were considered mainly to be supply-side driven.

3.3 FINANCIAL FACTORS

In addition to commodity attributes, crude oil also has financial attributes. The long term trend of crude oil price is determined by the commodity attributes, which are affected by the supply and demand factors generated by the real economy; the short term fluctuations of crude oil price are determined by the financial attributes.

Which are influenced by market expectations and speculative transactions. The financial factor mainly includes speculation factor, exchange rate and some other financial index

Which connect the stock market and monetary market with the crude oil price development of a structural model to estimate the speculative component of oil price through the inventory data and found it played an important role during earlier oil price shock period.It estimated the comovement and information transmission among oil price, exchange rate and the spot prices of four precious metals (gold, silver, platinum, and palladium).

Investors could diversify their investment risk by investing in precious metals, oil, and euros.

3.3 FINANCIAL FACTOR

The Crack spread is defined as the price difference between crude oil and its refined oil, reflecting the supply and demand relationship between the crude oil market and its refined product market.Many used the random walk model(RWM) as a benchmark to compare the crack spread futures and crude oil futures and found the crack future could forecast the movements of oil spot price as reasonable as the crude oil futures.

Then they selected crack spread as one of the variables to forecast crude oil prices, and the studies suggested it was an influential prediction factor.

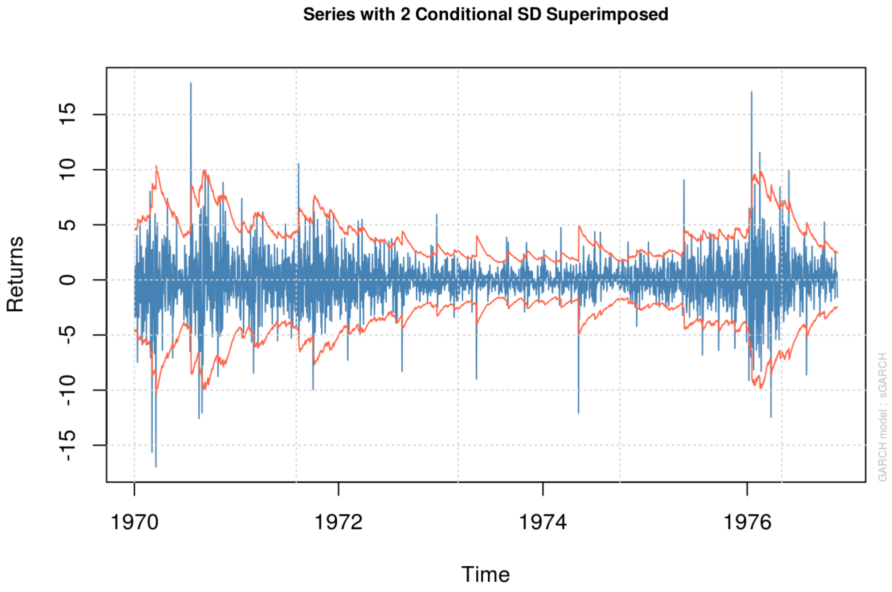
3.4 FORECAST METHOD

Except for the influence factors, researchers are also very concerned about the forecast methods for improving forecast accuracy. The four main forecast method categories: time series models, econometric models, qualitative methods and artificial intelligence

4.PROJECT THEME

Crude oil price market prediction is known for its obscurity and complexity. Due to its high vacillation degree, unpredictable irregularity events, and the complex correlations involved between the factors in the market, it is indeed difficult to predict the movements of the crude oil price. The crude oil market has strong evidence of chaos and develops as one of the most volatile markets in the world. Corresponding to that, there are few numbers of research conducted for crude oil price prediction.

Fig 4.1 Generalized Autoregressive Conditional Heteroskedasticity (GARCH)



Among the research models used are the single statistical and econometric model, single Artificial Intelligence (AI) model and the hybrid. Formerly, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model and Naive Random Walk were among the statistical and econometric models used to predict crude oil price. Research successfully utilized a probabilistic model to predict the oil price. The research was conducted based on a case study about the probabilistic inheritance of RNN models.

The models are used to forecast crude oil price and then produce a probabilistic prediction for it . The probabilistic prediction is actually generated by running Monte Carlo analysis on annual WTI average prices. For the purpose of the simulation experiment , the analysis done in this study is based on two assumptions of the timing when Iraq's return to the market and the impact of oil exports from the Former Soviet Union.

Three variables input are then used to define the scenarios; the probabilities of embargo ends, total demand and other world productions. The results from the simulation were robust and consistent with the annual average prices are almost certain between a range of amounts per barrel.

There was only 0.75% out of the total scenarios, predicted price over the range. Other statistical model predictions made for crude oil price are by C. Morana . This research used a semi parametric approach suggested for short-term oil price prediction. It also used GARCH to employ oil price changes to predict the oil price distribution over the short-term horizon.

The approach used one-month-ahead daily Brent oil price which emphasized periods with high uncertainty during some period. Furthermore, the analysis of the forecasting is based on the last two months of the available data and according to the analysis the result was strayed from the actual. This is most likely linked to the widening of the forecast confidence interval.

Nevertheless, the study offers improvement from the next statistical model used for predicting the crude oil market is by where they predict monthly WTI spot price using relative inventories. This study used Relative Stock Model (RSTK) as the basis to predict the price by comparing two other alternative models Naïve Autoregressive (NAIV) forecast model and Modified Alternative (MALT) model.

The only variable they used in this research is the petroleum inventories because of its independent practicality and it is readily available every month. The RSTK model shows the best performance for both in and out of sample forecast compared to the other two models. It is also being used by the Energy Information Administration (EIA) with among other models to investigate the future market disruptions that derived from changes in demand and production.

Nowadays, AI models are among the popular tools to be used for prediction. As an alternative tool to statistical and econometric models, AI offers recognition ability on complex patterns and also on providing intelligent reasoning and intelligent decision-making based on data. Among the single AI models used for predicting the crude oil price is Support Vector Machine (SVM) in which for the task of time-series prediction, this research focused only on the Support Vector Regression (SVR) model.

5.WORKING PRINCIPLE

The user needs to access the data model built by us. In that model we have added predictions and inputs are given to it which mainly deals with crude oil price related data. Then for the model we do evaluation which is a process of finding whether the model is enough for proper evaluation and consistent use.

Initially we have a dataset that deals with crude oil price related data.e term data set refers to a file that contains one or more records. For that data we do data processing. Data processing is essential for our project to create better ideas about our project. By converting the data into readable formats like graphs, charts, and documents, employees throughout the organization can understand and use the data.

Then after data processing we do training and testing is an extremely large dataset that is used to teach a machine learning model. Training data is used to teach prediction models that use machine learning algorithms how to extract features that are relevant to specific business goals. For supervised ML models, the training data is labeled. The data used to train unsupervised ML models is not labeled.

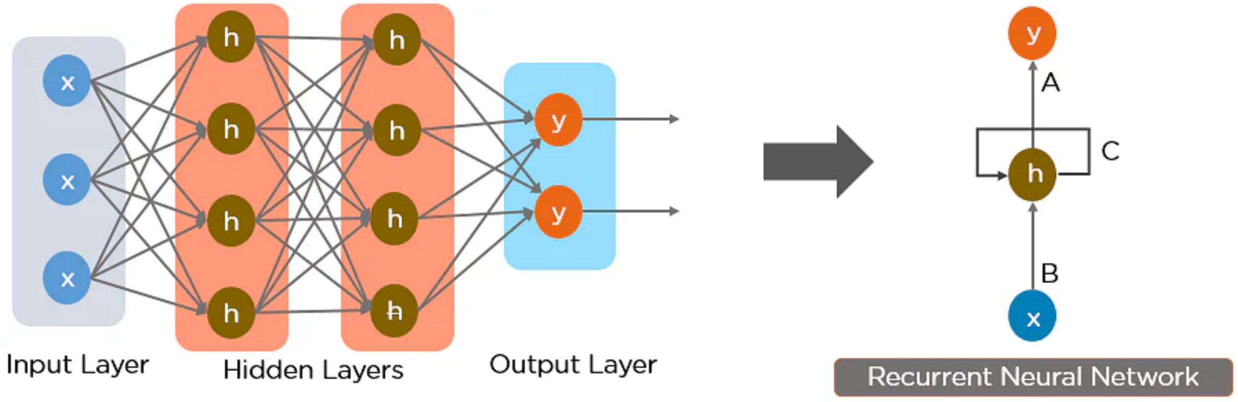
The idea of using training data in machine learning programs is a simple concept, but it is also very foundational to the way that these technologies work. The training data is an initial set of data used to help a program understand how to apply technologies like neural networks to learn and produce sophisticated results. It may be complemented by subsequent sets of data called validation and testing sets

Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another subset of original data, which is independent of the training dataset.

However, it has some similar types of features and class probability distribution and uses it as a benchmark for model evaluation once the model training is completed. Test data is a well-organized dataset that contains data for each type of scenario for a given problem that the model would be facing when used in the real world. Usually, the test dataset is approximately 20-25% of the total original data for an ML project.

At this stage, we can also check and compare the testing accuracy with the training accuracy, which means how accurate our model is with the test dataset against the training dataset. If the accuracy of the model on training data is greater than that on testing data, then the model is said to have overfitting.

We apply the RNN algorithm to the model Recurrent Neural Network(RNN) is a type of Neural Network where the output from the previous step is fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words. Thus RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is Hidden state, which remembers some information about a segue.

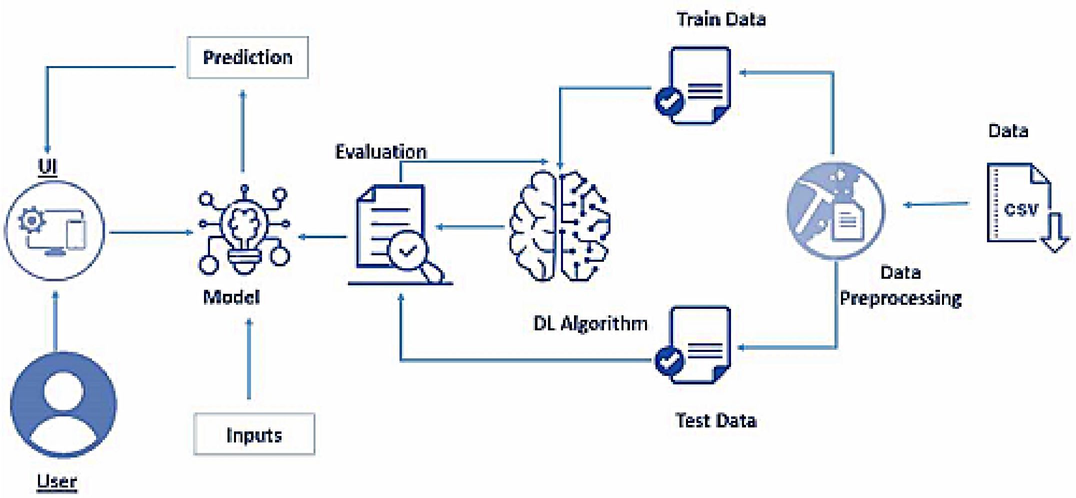


The input layer ‘x’ takes in the input to the neural network and processes it and passes it onto the middle layer.

The middle layer ‘h’ can consist of multiple hidden layers, each with its own activation functions and weights and biases. If you have a neural network where the various parameters of different hidden layers are not affected by the previous layer, ie: the neural network does not have memory, then you can use a recurrent neural network.

The Recurrent Neural Network will standardize the different activation functions and weights and biases so that each hidden layer has the same parameters. Then, instead of creating multiple hidden layers, it will create one and loop over it as many times as required.

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. It is used for processing, predicting, and classifying on the basis of time-series data.



After applying the required data learning algorithm we do an evaluation of the model and it is ready for the user to run and give data such that we are able to predict the data .

6.IMPLEMENTATION SCREENSHOTS

7.CONCLUSION

We have developed a model that Oil demand is inelastic, therefore the rise in price is good news for producers because they will see an increase in their revenue. Oil importers, however, will experience increased costs of purchasing oil. Because oil is the largest traded commodity, the effects are quite significant. A rising oil price can even shift economic/political power from oil importers to oil exporters. The crude oil price movements are subject to diverse influencing factors.

Crude oil is one of the most powerful types of energy and the fluctuation of its price influences the global economy. Therefore, building a scientific model to accurately predict the price of crude oil is significant for investors, governments and researchers. However, the nonlinearity and nonstationarity of crude oil prices make it a challenging task for forecasting time series accurately. To handle the issue, proposed model has forecasting approach for crude oil prices that combines Recurrent Neural network Long Short-Term Memory (LSTM) with attention mechanism and addition, following the well-known “decomposition and ensemble” framework

This Project mainly focuses on applying Neural Networks to predict the Crude Oil Price.This decision helps us to buy crude oil at the proper time. Time series analysis is the best option for this kind of prediction because we are using the Previous history of crude oil prices to predict future crude oil. So we would be implementing RNN(Recurrent Neural Network) with

LSTM(Long Short Term Memory) to achieve the task.

8.REFERENCES

- Galyfianakis G., Garefalakis A., Mantalis G. (2017) The effects of commodities and financial markets on crude oil, Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles 72, 1, 3. [Google Scholar]
- Wang Y., Wei Y., Wu C. (2011) Detrended fluctuation analysis on spot and futures markets of West Texas Intermediate crude oil, Phys. A, Stat. Mech. Appl. 390, 5, 864–875. [Google Scholar]
- Abledu G.K., Agbodah K. (2012) Stochastic forecasting and modelling of volatility of oil prices in Ghana using ARIMA time series model, Eur. J. Bus. Manag. 4, 16, 122–131. [Google Scholar]
- Baumeister C., Kilian L. (2012) Real-time forecasts of the real price of oil, J. Bus. Econ. Stat. 30, 2, 326–336. [Google Scholar]
- Shin H., Hou T., Park K., Park C.K., Choi S. (2013) Prediction of movement direction in crude oil prices based on semi-supervised learning, Decis. Support Syst. 55, 1, 348–358. [Google Scholar]
- Yu L., Xu H., Tang L. (2017) LSSVR ensemble learning with uncertain parameters for crude oil price forecasting, Appl. Soft Comput. 56, 692–701. [Google Scholar]

- Tang M., Zhang J. (2012) A multiple adaptive wavelet recurrent neural network model to analyze crude oil prices, J. Bus. Econ. 64, 4, 275–286. [Google Scholar]
- Zhao Y., Li J., Yu L. (2017) A deep learning ensemble approach for crude oil price forecasting, Energy Econ. 66, 9–16. [Google Scholar]
- Kristjanpoller W., Minutolo M.C. (2016) Forecasting volatility of oil price using an artificial neural network GARCH model, Expert Syst. Appl. 65, 233–241. [Google Scholar]
- Safari A., Davallou M. (2018) Oil price forecasting using a hybrid model, Energy 148, 49–58. [Google Scholar]
- Yu L., Zhao Y., Tang L. (2014) A compressed sensing based AI learning paradigm for crude oil price forecasting, Energy Econ. 46, 236–245. [Google Scholar]
- Zhang X., Lai K.K., Wang S. (2008) A new approach for crude oil price analysis based on empirical mode decomposition, Energy Econ. 30, 905–918. [Google Scholar]
- Xing F.Z., Cambria E., Welsch R.E. (2018) Natural language based financial forecasting: a survey, Artif. Intell. Rev. 50, 1, 49–73. [Google Scholar]