Big Data Analytics and Machine Learning for Crude Oil Price Prediction

Jiahe Zhou, Xuyang Xiao, Chengnan Xu, Tian Yu

Abstract

Crude oil forecasting has become increasingly prevalent globally, due to the economic value attached to the product of crude oil such as gasoline, diesel, heating oil, jet fuel, etc. Consequently, there is no denying that one of the most important roles of economic variables in today's world countries are the price and the change of crude oil price. The variation in the price of crude oil is critical to the strategy of budget and treasury. Thus, accurate forecasting of the crude oil price and realization of the forecasts based on this forecast will provide significant information for savings or gains in government and corporate economics concerning with some policy and planning issues. In this paper, we compared different machine learning algorithms, such as the Support Vector Regression, Random Forest and Gaussian Process, analyzed the prediction results of those methods and found the optimal algorithm and the best fitted model.

Introduction

Crude oil, which is known globally as ‘Black Gold’, plays an indispensable role in the world’s economy, especially in the few several decades. Commonly, crude oil is used in a wide range of fields including industry, transportation, automobile, cosmetics, energy and oil had become a turning point in terms of country economics.

Admittedly, crude oil prices are highly sensitive to political and economic developments. For instance, the increasing geopolitical risks in the Middle East cause high volatility in oil prices. Meanwhile, the prices of crude oil can be affected by the following factors that must be considered.

- Oil Reserve Amount
- Development in economics
- Global climate changes
- Technique inventions
- Changes in supply and demand balance
- The increase or decrease in demand among interrelated countries.

Additionally, forecasting crude oil prices is important as it affects other key factors of the economy including the consumption industry, electricity and power industry, vehicles market and the stock market.

In recent years, the crude oil market have witnessed an extreme fluctuations in the price of international crude oil, as is shown below the price index monthly price for crude oil during the last 20 years, the numbers are the calculation of the average between the prices from Dated Brent, Dubai Fateh and West Texas Intermediate. we can see that due to the world-wide financial crisis, the crude oil market suffered a dramatic increase as well as the decrease between the year 2007 and 2009. Hence, it is critical for the government to monitor the monthly or even daily price, supply and demand of crude oil and provide the timely changes in the import or export of crude oil. Undoubtedly, the changes in price can not only affect the financial markets and

economics, but it can also exert an effect on individuals due to the fact that plenty of goods and services are produced by crude oil, which eventually influence the gross domestic product(GDP).

Figure 1. Crude Oil (petroleum), Price Index Monthly Price - Index Number

Due to the corresponding reasons that mentioned above, the importance of conducting such a model can be attached in the following aspects:

- Decrease the impact of price fluctuations
- Assist government and policy makers to take actions

In order to give an estimation of the future price trend, firstly, the time series model can be adopted to provide a rough prediction of the price, followed by the modified SVR, Random Forest, Gaussian Process model to estimate precisely.

In machine learning, Support Vector Regression, Random Forest and Gaussian Process are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Based on that, we are trying to set up a modified model in a gesture to anticipate the future price trend.

Background

In contemporary society, with the rapid development of technology, there is no denying the fact that the last several decades have witnessed a dramatic development in terms of worldwide economy, under the condition that the world's unsustainable energy has a limited amount of supply, the competence over crude oil has becoming increasingly prevalent in the 21st century, for those countries where crude oil production is low and heavily dependent on crude oil import, this is of utter importance to have an estimation of the crude oil up to date. Just taking the United States of America as an example, based on the statics provided by *U.S Energy Information Administration(eia)* which is shown as Table 1 below.

Import Sources	Gross Import (MM barrels/day)	Exports (MM barrels/day)	Net Imports (MM barrels/day)
Total, all countries	8.01	1.15	6.86
Canada	3.20(40%)	0.157	3.04
Saudi Arabia	0.76(9%)	<0.01	0.75
Mexico	0.54(7%)	0.19	0.34
Venezuela	0.53(7%)	0.011	0.52
Iraq	0.74(6%)	<0.01	0.73

Table 1. Top Sources and Amounts of U.S Crude Oil Imports, 2017

From the above Table 1, it is convincing that U.S is the kind of countries that rely mainly on the imports of crude oil and produced far less amounts to exports. Under the condition that U.S the biggest economic entity, the prediction of crude oil price can be beneficial to the relevant government and organizations concerning with making policy options and import plans.

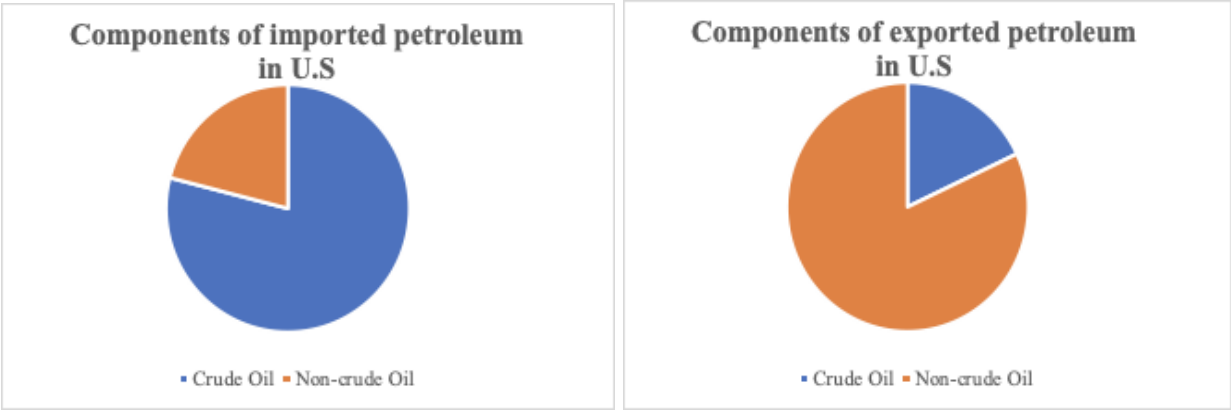


Figure 2. The Components of Imported and Exported Petroleum in U.S, 2017

Based on the above Figure 2, it is obvious that the crude oil consists of a large proportion of the imported petroleum in U.S compared with the non-crude oil including hydrocarbon gas liquids, refined petroleum products such as gasoline and diesel fuel, and biofuels including ethanol and biodiesel. Conversely, the proportion of crude oil in exported petroleum shows that the majority is non-crude oil which consists of almost 82%.

Recently, in the weekly petroleum data conducted by *U.S Energy Information Administration(eia)* for the week ending March 15, 2019, shows that the crude oil refinery increased 178,000 barrels per day which results in the increase of gasoline production, distillate fuel production and the jet fuel production. Consequently, the increase in operation causes the decrease in the total inventories of crude oil in the U.S, which is 439.5 million barrels in total. Meanwhile, due to the fluctuation of the price, supply and demand, also the volatility over the production of crude oil in the U.S country, crude oil imports decreased by 11.3% in 6.6 million barrels per day. Besides the U.S, there are also many countries with high demand of crude oil but low production, such as China, Japan, European Countries. Etc.

According to the previous information, we would like to collect over 300 samples to predict the price trend of crude oil. The data collected from *Index Mundi* gives us the current price and change rate worldwide from Feb, 1999 to Feb, 2019, shown as the average between the spot price from Dated Brent, Dubai Fateh and West Texas Intermediate.

Jul-18	72.67	0.96%
Aug-18	71.08	-2.19%
Sep-18	75.36	6.02%
Oct-18	76.73	1.82%
Nov-18	62.32	-18.78%
Dec-18	53.96	-13.41%
Jan-19	56.58	4.86%
Feb-19	61.13	8.04%

Figure 3. The Collected Data Sample

And the figure 4 gives us the price trend processed by the data shown in figure 3.

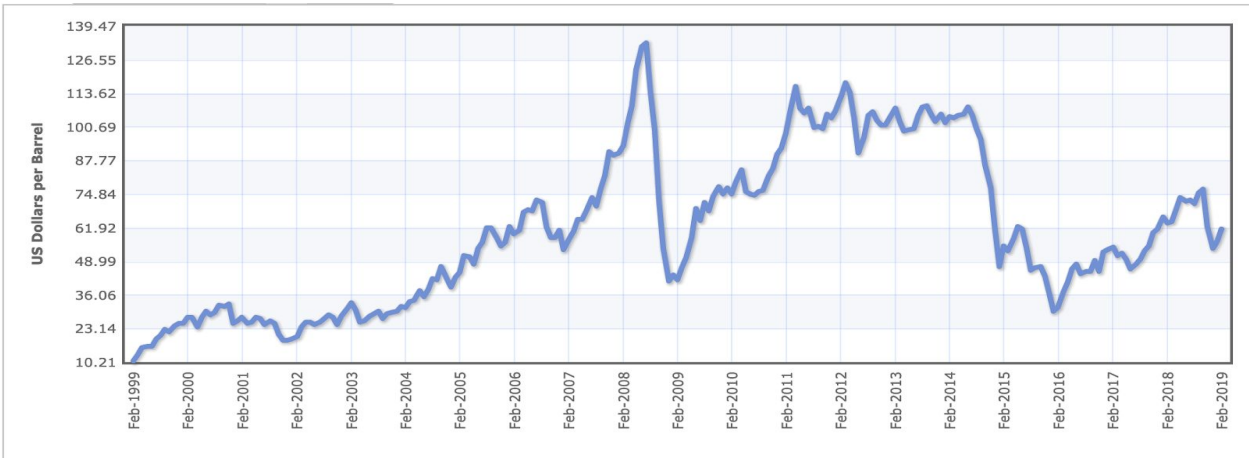


Figure 4. The Crude Oil Price (US Dollars per Barrel) between 1999-2019

The corresponding dataset we collected is mainly from the *Index Mundi*, *eia* and *bp*, for this project, we specifically considered six features based on the original dataset we collected from the websites, and they are:

- Date
- Seasonality
- Global production
- Global consumption

- Previous price
- Important events

Date	Seasonality	Global Production	Global Consumption	Previous Price	Important Event	Price
…	…	…	…	…	…	…
20180501	1	99536	99229	68.79	1	73.43
20180601	1	100307	100967	73.43	1	71.98
20180701	1	101185	100858	71.98	1	72.67
20180801	1	101542	101196	72.67	1	71.08
20180901	1	101518	99664	71.08	1	75.36
20181001	0	102586	100254	75.36	1	76.73
20181101	0	102201	100604	76.73	0	62.32
20181201	0	101696	100523	62.32	0	53.96
20190101	0	100031	99294	53.96	0	56.58
20190201	0	100314	101812	56.58	0	61.13

Table 2. The Original Dataset

For the date feature, since we collected the dataset by month, so we assign each month with a specific number, for instance, we set Mar. 1989 as our starting month 1, and Feb. 2019 as our ending month 360. Besides, after we looked into the original dataset, there is an obvious seasonality in every single year, so we set the hot months when the crude oil price is higher to 1 and the cold months when the crude oil price is lower to 0. The global production and consumption in our dataset represents the crude oil market condition which can have a direct influence on the crude oil price. As for the previous month price feature, since the crude oil is a future goods, its most recent month price can directly influencing the crude oil market.
Another indirect feature is the important event. Actually, most of political & economical events can also affect the crude oil price in various aspects. From Table 2&3, we can see that there are many historical events in the timeline correlating to the fluctuation of the crude oil price, which are the Gulf War from August 1990 to February 1991, the Asian financial crisis in July 1997 , the September 11 attacks on the morning of September 11, 2001, the War in Afghanistan starting from October 2001, the Iraq War in 2003, the United States subprime mortgage crisis occurring between 2007 and 2010, the global financial crisis of 2007-2008, the Arab Spring in late 2010,

the Iran nuclear deal in July, 2015, the trade war between China and the US starting from January 22, 2018 etc.

Figure 5. Historical events influencing crude oil price

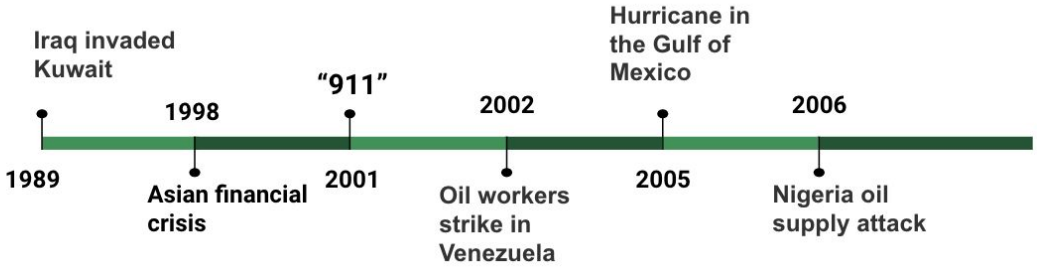


Figure 6. The timeline of the historical events

Before we implement the machine learning algorithm, we normalize the variables into scale 0 to 1. We use minmaxscaler package in Python to implement the normalization. Then, we set the originally 270 data as the training data while the rest 90 ninety data as the test data.

Month Number	Seasonality	Global Production	Global Consumption	Previous Price	Important Events	Price
...	...	...	...	...	...	...
351	1	0.92138	0.92833	0.47688	1	73.43
352	1	0.94125	0.97655	0.51479	1	71.98
353	1	0.96389	0.97353	0.50294	1	72.67
354	1	0.97309	0.98291	0.50858	1	71.08
355	1	0.97247	0.94040	0.49559	1	75.36
356	0	1.00000	0.95677	0.53055	1	76.73
357	0	0.99008	0.96648	0.54174	0	62.32
358	0	0.97706	0.96423	0.42403	0	53.96
359	0	0.93414	0.93013	0.35574	0	56.58
360	0	0.94143	1.00000	0.37714	0	61.13

Table 3. Dataset Normalization

Literature review

Changes in the price of crude oil have a very critical role in terms of treasury and budget, both in company and state planning. Accurate forecasting of the crude oil price and realization of the forecasts based on this forecast will provide savings or gains in government and corporate economies. Therefore, there is a great need for this estimation in countries which heavily dependent on crude oil importing. Mohammad Reza Mahdiani, Ehsan Khamehchi [1] modified the neural network model to predict the daily and monthly crude oil prices very accurately. They compared the model with the pure neural network and confirmed the model's greater performance in situation of small number of input data for training or the great changes of variables. Mesut Cumus and Mustafa S. Krian[2] interpreted the parameters affecting the crude oil prices by using gradient boosting algorithm and control the bias-variance trade-offs in the estimation. With the help of component wise gradient boosting, they can add automatic variable selection during the fitting process.Adnan Khashman, Nnamda I. Nwulu[3] present a support vector machines system that predicts the price of crude oil with a high degree of accuracy in predicting crude oil price. They totally have 1252 observations from USA- EIA website where the last 626 observations used for training, while the first 626 observations were used for testing the trained prediction neural model.

In 2014, Rana Abdullah Ahmed, Ani Bin Shabri[4] present a novel technique for forecasting technique for forecasting crude oil price based on Support Vector Machines(SVM) and employs ARIMA and GARCH method to evaluate the performance of the proposed model.
Ardalan Tebyanian, Fares Hedayati[5] established two ensemble regression algorithm for forecasting the daily price of crude oil form features extracted from the U.S. Energy Administration and some international news agencies.

With prediction techniques developing, more researchers devort themselves into this study. Yang Zhao, Jianping Li, Lean Yu[6] proposed deep learning ensemble approach to capture the behavior of crude oil price. The new approach combines the merits of the stacked denoising autoencoders technique and bootstrap aggregation method and is especially suitable for oil price forecasting. Wen Xie, Lean Yu, Shanying Xu etc.[7] implemented the procedure of developing a support vector machine model for time series forecasting involves data sampling, sample preprocessing, training & learning and out-of-sample forecasting demonstrated the excellent performance by comparing with those of ARIMA and BRNN.

Methods

The price of crude oil can be viewed as highly nonlinear and to some extent, the data can be seen completely random, so the traditional time series model including the ARIMA, GARCH may not be effective for all time. In this paper, we will adopt Support Vector Regression, Random Forest and Gaussian Process model to predict the possible price trend for crude oil. First of all, the implementation of traditional time series models which is the Simple Moving Average and Exponential Moving Average(EMA) will be tested under the preprocessed dataset in order to give a rough estimation of the price and provide comparison with the machine learning models.

1 Time Series Models

1.1 Simple Moving Average(SMA)

The Simple Moving Average[8] calculates the average price of the original n data samples and assign it to the next data sample and for each time, take off the first one and include the recent data sample. The formula for SMA is as follows:

$$SMA = \ (A1 + A2 + A3 + ... + An)/N$$

Where:

- N = *The number of months in the model period.*
- An = *The crude oil price of the* n^{th} *month.*

1.2 Exponential Moving Average(EMA)

The Exponential Moving Average[9] follows a similar process as the SMA method, by introducing a multiplier in order to weighting the EMA. The formula for EMA is as follows:

$$EMA = Price(t) * k + EMA(y) * (1 - k)$$

Where:

- N = *The number of days in EMA which can be obtained by the tester*
- $k = 2 \div (N + 1)$

2 Machine Learning Algorithms

2.1 Support Vector Regression(SVR)

SVM[10] is commonly adopted in pattern classification and nonlinear regression. Support Vectors have the advantage of reducing the problems of over-fitting or local minima since it is based on the structural risk minimization principle..

The SVM algorithm seeks one misalignment mapping from the input space to output space ϕ. Through this mapping, data X is mapped to a feature space Γ, and linear regression is carried out in the feature space with the following function: The SVM regression equation is as follows:

$$f(x) = [w \times \varphi(x)] + b$$

$$\phi : R^m \rightarrow \Gamma$$

In above, b is a threshold value. According to statistical learning theory, SVM determines the regression function through objective function minimization:

$$\min\{0.5\, w^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)\}$$

S.t. $yi - w^* \varphi(x) - b <= \varepsilon + \xi_i^*$

$w^* \varphi(x) + b - yi <= \varepsilon + \xi_i$

$$\xi_i,\ \xi_i^* > 0$$

The SVM regression equation is as follows:

$$f(x) = \sum_{i=1}^{n} (ai - ai^*)\, K(Xi, X) + b$$

where $K(X_i, X)$ is the SVM kernel function. Kernel function types include linear kernels, polynomial kernels, and radial basis function Kernel(RBF), and the Sigmoid Kernel.

2.2 Random Forest

Random forest[11] is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or mean prediction of the individual trees. Random forests is a trademark term for an ensemble of decision trees, that it can be used for both regression problems and classification. Random forests does not overfit and can run as many trees as you want.

Also it is very fast, for example, running on a data set with 50,000 cases and 100 variables, it produced 100 trees in 11 minutes on an 800 Mhz machine. Therefore, in our small data set, it will work faster. In addition, random forest is very user-friendly since it only has two parameters (the number of variables in the subset and the number of trees in the forest) and therefore the values is not really sensitive[12].

2.3 Gaussian Process

Gaussian Process (GP) are a generic supervised learning method designed to solve regression and probabilistic classification problems. The Gaussian Process which is a kind of algorithm fit

small sample data, different kernels can be specified and provide common kernels. And also it can avoid the problem of overfitting. Rather than deal with complicated process such as prior on function spaces, only we considered is that the text point x, the training data like x1, x2,.....,xn and the test points, and then the value to we wish to predict[13].

Since a Gaussian process is a generalization of the Gaussian distribution and is entirely specified by a mean and covariance functions. Then we need only concerns about covariance matrices, namely the finite-dimensional objects.

In the gaussian process regression, the key point is to find a suitable kernel. (it is also called "covariance functions"), which is kind of crucial ingredient to help determine the shape of prior and posterior. Therefore various kernels were tried to improve the fit of the model and validation were done using the MSE and R-square. Two kinds of categories of kernels can be applied: stationary kernels depend only on two data points and not on their absolute values K(xi,xj)=K(d(xi,xj)), while non-stationary kernels depend also on data points.

3 Validation Methods

3.1 Mean Square Error

The Mean Square Error measures the average of the squares of the errors, which is the average squared difference between the estimated values and what is estimated. Commonly, it is a risk function, corresponding to the expected value of the squared error loss. In this project, by compared with different methods, we are trying to find the best model with the smallest MSE. And the formula is given below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

3.2 Root Mean Square Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are also method of measuring the differences between predicted and real values. RMSE can help figure out how spread out these residuals are. It can tell people that data

concentrated enough or not to help the line of best fit. RMSE is kind of method to help verify the results in regression analysis. By comparing different RMSE, we need to find the optimal solutions with minimum RMSE.

The formula is given below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}}$$

Where:

Yi=observed values.(known results)

$\hat{Y}_i$ =forecasts(expected values or unknown results)

N=sample size

3.3 R-Square

R-squared is also kind of seemingly intuitive measure way to help calculate and compare how close the data between fitted regression line and the real data. It is the percentage of response variable variation that is explained by the regression model. The higher value of the R-square represent a better performance of the algorithm.

The formula is given below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

3.4 Cross Validation

In machine learning method, there is a risk of overfitting, which means that sometimes the data would just perfect score but fail to predict the unseen data. To solve this problem, we use the cross validation (CV) to help avoid such problems. The training set is split into k smaller sets and k-1 folds used as training data. The resulting model is validated on the remaining part of the data, used as the test set to help measure the accuracy. In our model, in SVR, we use cross validation to help measure the accuracy. And we use python, to call cross_val_score function to help estimate the dataset.

The flowchart of cross validation shown as below:

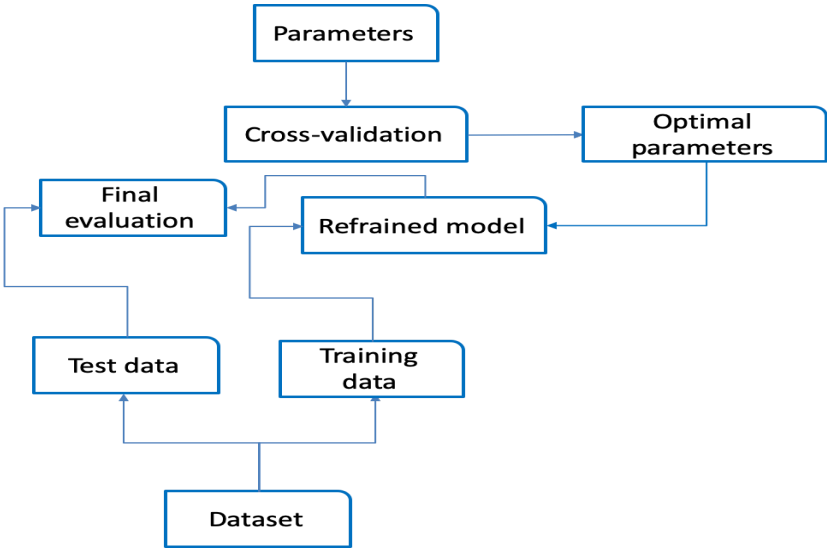


Figure 7. The Cross validation method

Results and Discussion

1. Time Series Models

For the time series models which are the Simple Moving Average and Exponential Moving Average, for the SMA, we considered that the time period is 12 months which equals to one whole year, and there are several reasons:

- The shorter time period can not take some seasonality or price trend in the single year into account.
- The longer time period can reduce the influence of some high price month due to some political issues or world events and reduce the accuracy of the prediction.

Compared with the SMA model, in the EMA model, we attached the importance about the previous month price which is shown as the ‘k’ in the model, under the fact that the previous month can have more impact on the following month than the original several months, so we put the weights into the model. And the following figure shows the prediction results compared between EMA, SMA and the original dataset.

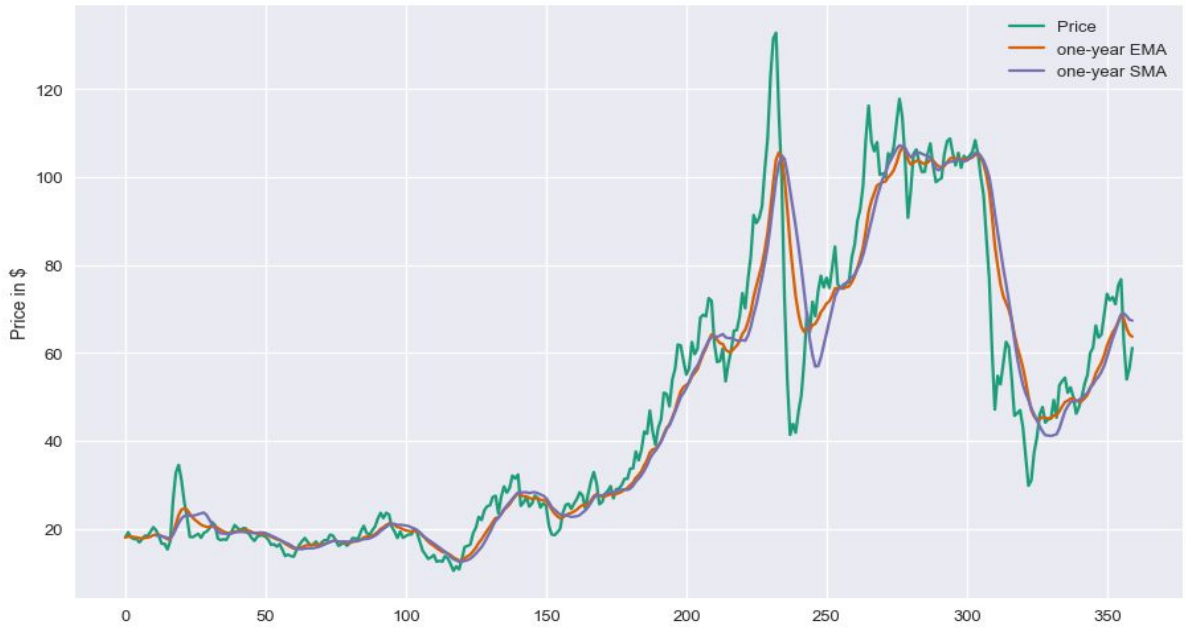


Figure 8. The prediction results of EMA and SMA

From the above we can see that both the SMA and EMA model can provide a good fit in those data samples. Conversely, both the time series models failed to explain the extremely high price months. And the results for this time series model is as following:

	MSE	**RMSE**	**R-Square**
SMA	124.87	11.174	0.8152
EMA	95.75	9.785	0.8583

Table 4. Time series models prediction results

Based on the output results of the time series models, we can assume that this kind of statistical methods can have a good prediction of the crude oil price, that is to say, the R-Square, which represents the performance of the prediction models, is high enough to have a good prediction of the crude oil price. However, since the time series model can only give a prediction of the following one month, it is difficult for this kind of models to predict a trend of the following

several months' price. As we can see from the above table, the mean square error for both the models is around or over 100, even though we do not scale the price index into 0 to 1, these are still very large numbers.

2. Machine Learning Algorithms

As is mentioned above, although the time series models are widely used in stock price prediction and many other statistical fields, it is still far away from providing an accurate price prediction for the future month, based on the fact that the crude oil price is essential to the stock market, the global economy and the individual strategy for each country, so we introduced the machine learning methods.

2.1 Data Preprocessing

In this project, we considered six features in the original dataset, which is the Year, Seasonality, Previous Month Price, Monthly Consumption, Monthly Production and the important events in the previous 30 years. Before we implemented the machine learning methods, it is essential to normalize the dataset, since the dataset we collected and the features we included in this dataset are highly related with the crude oil price, so the PCA method[14] was not used in this project to decompose the dataset. However, in order to keep all the input numbers having the same importance on the prediction model, we normalize the dataset into 0 to 1 by using the python package 'MinMaxScaler'. By doing this, we can give the six features normalized importance and influence on both the training data and the test data.

Month_level	**October--March**	**April--September**
Scale Number	0.0	1.0

Table 5. The input number of seasonality

2.2 Data Structure

For this project, we used the original 270 data samples to be the training data and the rest 90 data samples to be the test data, and the following figures show the original dataset split by the training data and test data.

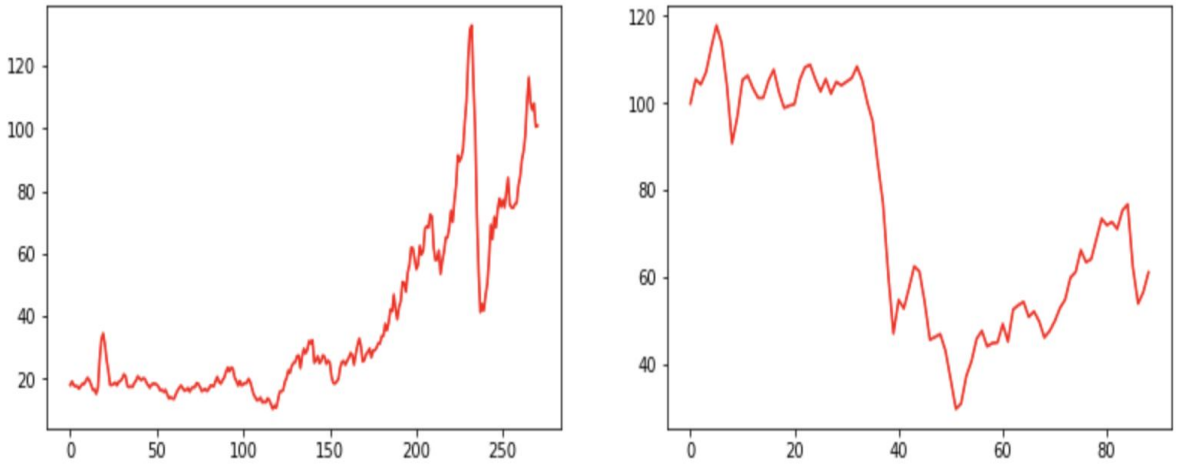


Figure 9. The Original world crude oil price

From the above figures, there is an obvious fluctuation of the crude oil price among the past 30 years. And that is also the reason why we choose the machine learning algorithms to have a better prediction for the future.

2.3 Support Vector Regression

The Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm which is the maximal margin. There are several advantages that the SVR algorithm has, for instance, SVR model tends to have strong generalization ability and can reduce the probability of overfitting. There are several kernels in this model that needed to consider, they are as follows:

- Linear Kernel
- Polynomial Kernel
- Radial-basis Function(RBF) Kernel
- Sigmoid Kernel

In the support vector regression model, the RBF kernel is also known as the Gaussian Kernel, after we compared those four different kernels, we found out that the RBF kernel is the best fit for our model. And the formula for this kernel function is:

$$K(x,y) = \exp(-\gamma \| x-y\|^2), \gamma > 0$$

In this kernel function, we need to focus on two different hyperparameters, which is:

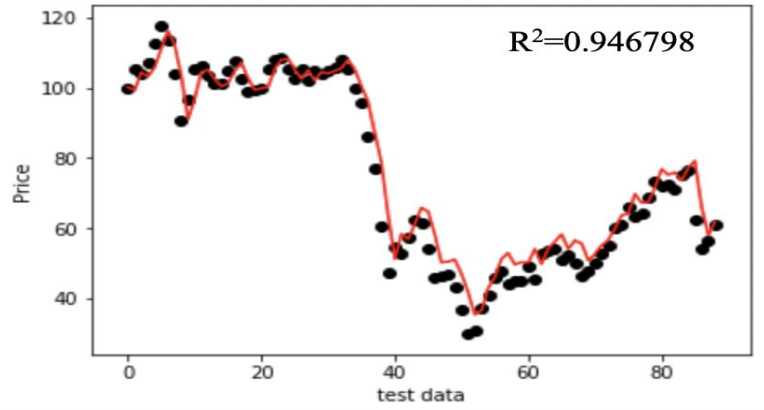
- C: Cost parameter of error item
- Gamma: The parameter in the kernel function

In this model, in a gesture to avoid overfitting, we used the Cross Validation with k-ford equals 5 to find the optimal hyperparameters in the Support Vector Regression model. And the optimal hyperparameters we got based on the input of our training data is:

Parameters	**C**	**Gamma**
Values	10,000	0.001

Table 6. The optimal parameters in RBF model

As is mentioned above, the optimal parameters we found is when C equals to 10,000 and gamma equals to 0.001. And the prediction results are following:



C=10000,gamma=0.001

Figure 10. The optimal prediction result in SVR

From the plot we can see that the SVR model gives a perfect fit of the training data and an excellent prediction for the test data, and the optimal R-Square we can get from this model is 0.946798 which is very close to the original test data samples. Under the fact that the hyperparameters can actually influence the fitting of the model, we can see from the below figures, showing that how do the hyperparameters change the prediction results.

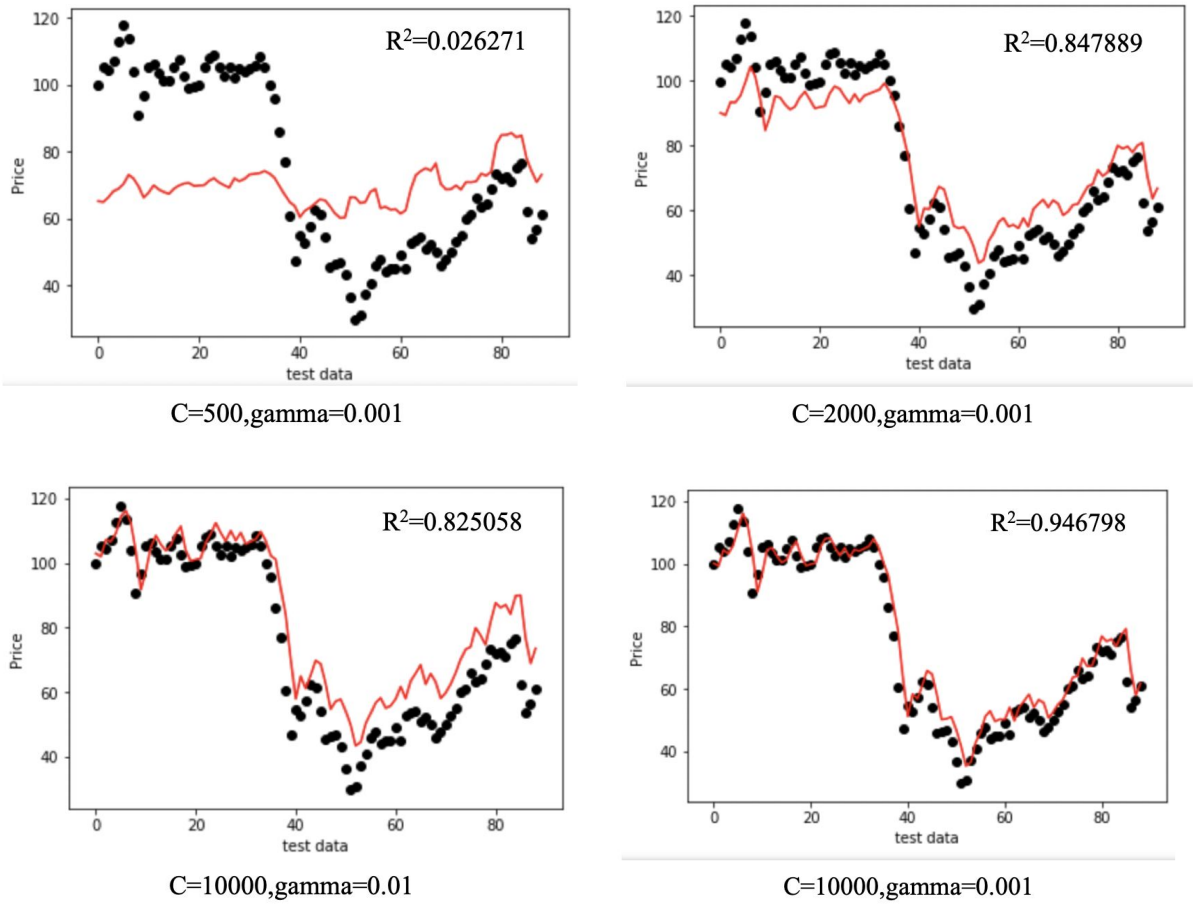


Figure 11. Different prediction results from SVR model

As we can easily get from the prediction results mentioned above, while the cost parameter of error item is very small, it can result in the condition of under-fitting, which means that the model fail to capture most of the data samples in the training data, and cause the prediction result is far away from the real world crude oil price. Once we increase the cost parameter, the prediction accuracy will increase at the same time. After we find out the optimal

hyperparameters, if we keep increasing the cost parameters, it will eventually result in the over-fitting of the model, which means the model has excessive pursuit of precision. And the validated results are shown below, compared with the time series model which is discussed previously, the SVR model can give a better prediction of the crude oil price as it has a smaller mean square error and larger R-Square.

	MSE	**RMSE**	**R-Square**
SVR(C=500, gamma=0.001)	658.24	25.656	0.0263
SVR(C=2000, gamma=0.001)	102.82	10.140	0.8479
SVR(C=10000, gamma=0.01)	118.25	10.874	0.8251
SVR(C=10000, gamma=0.001)	35.96	6.000	0.9468

Table 7. Different hyperparameters performance comparison

2.4 Random Forest

Random forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classification or mean prediction of the individual trees. There are several advantages[15] to this method:

- Small number of data loss does not change the regression model
- The high generalization ability
- Avoid the probability of overfitting
- Do not need to choose the hyperparameters specifically

For the random forest model, there are also some parameters that we need to consider:

- Number of features: the quantity of features in the model
- N estimators: the performance of the model
- Min sample leaf: the ability of data capturing
- Criterion: the validation criterion

While we implement the random forest model, we take the following values for our regression model to find the optimal prediction result. For the estimators, with the higher number of estimators can result in better performance in regression and prediction, but simultaneously, it can also bring the fact that the processing time will be much longer.

Parameters	**Features**	**Estimators**	**min_sample_leaf**	**Criterion**
Values	6	90	2	mse

Table 8. Optimal hyperparameters value

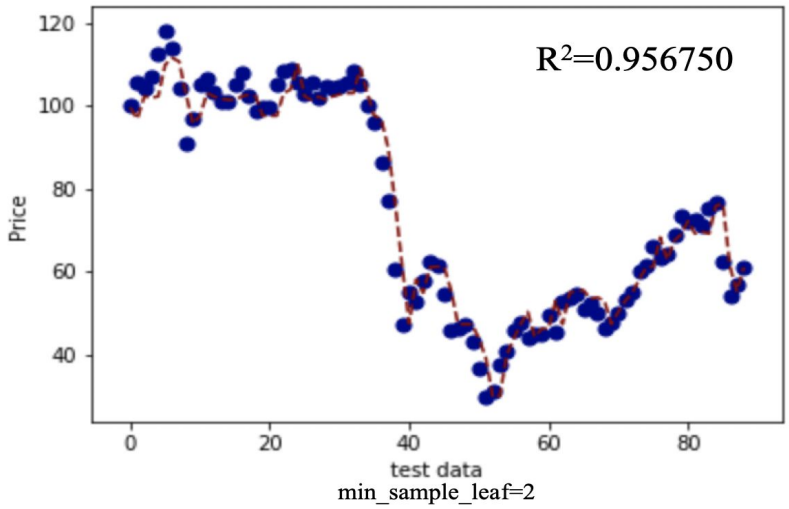


Figure 12. Optimal prediction results for random forest model

From the plot we can see that the Random Forest model also gives a well-performed fit of the training data and an excellent prediction for the test data, and the optimal R-Square we can get

from this model is 0.956750 which is very close to the original test data samples. Under the fact that the hyperparameters can actually influence the fitting of the model, we can see from the below figures, showing that how does the min_sample_leaf change the prediction results.

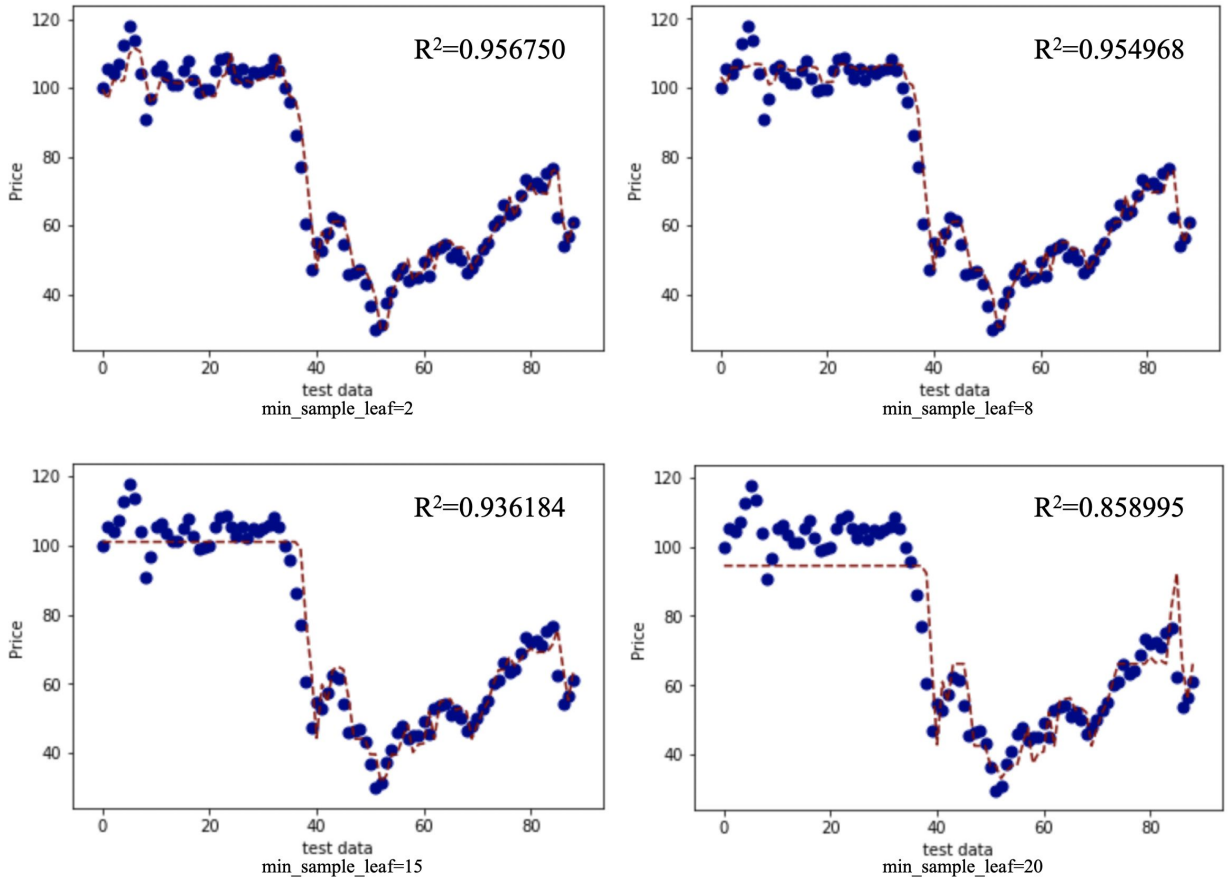


Figure 13. Prediction results based on the changes of min sample leaf

	MSE	RMSE	R-Square
RF(min=20)	95.31	9.763	0.8590
RF(min=15)	43.14	6.568	0.9362
RF(min=8)	30.44	5.517	0.9550
RF(min=2)	29.24	5.407	0.9568

Table 9. Different hyperparameters performance comparison

Based on the above prediction results, we can see that the min sample leaf represents the ability of data capturing, which means that with a larger value of min sample leaf, the model failed to capture the extremely high monthly price and tended to capture those price around the average monthly crude oil price which finally resulted in the underfitting of the model. As is shown in the above figures, that is, with a very large min sample leaf can eventually fail to capture a bunch of data samples. Conversely, the smaller values of this parameter can lead to overfitting of the model. Besides, compared with the SVR model which is mentioned above, we can see that the random forest model can provide a better prediction accuracy.

2.5 Gaussian Process

For gaussian process regression, since it can fit any black-box function and capture uncertainty. Different kernels can be specified and therefore, several kernels were tried to improve the fit of the models. After comparing with different kernels and find that the DotProduct kernel is the optimal one. Below we can see several different kernels in gaussian process regression.

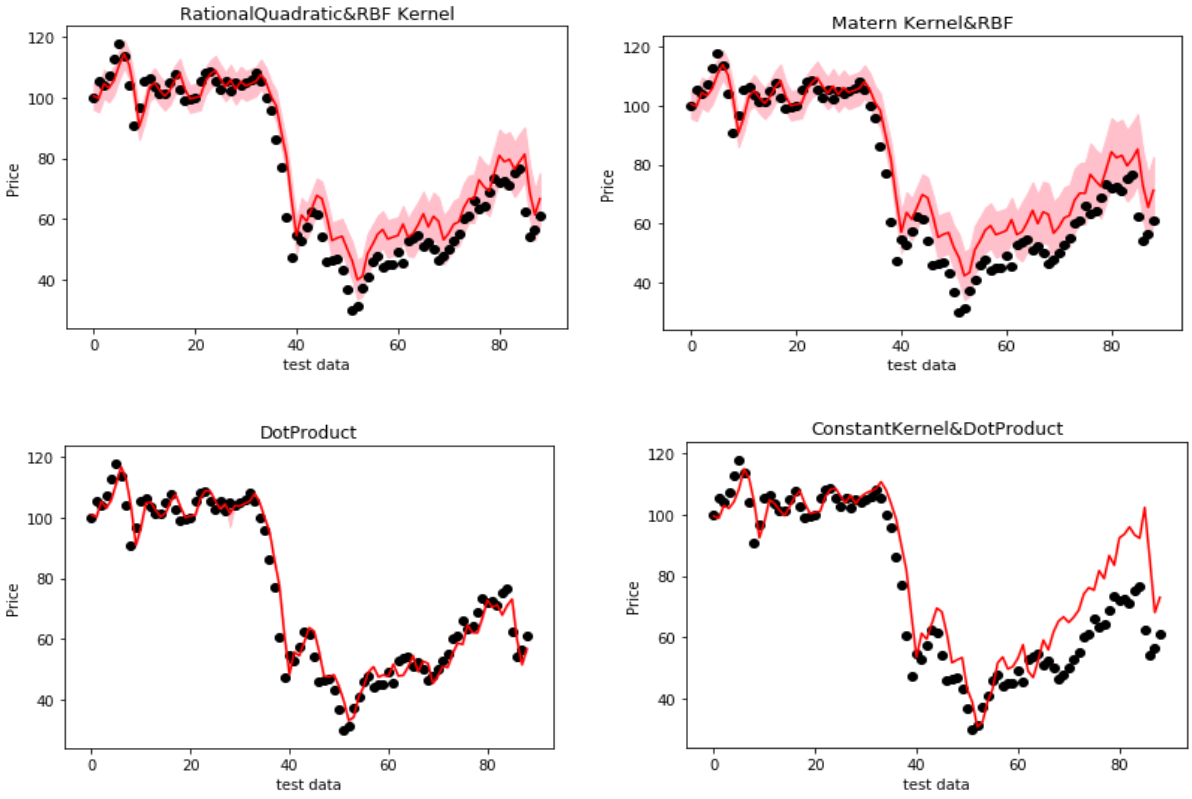


Figure 14. Prediction results based on the changes of different kinds of kernels

Since in Gaussian process regression, for different kernels, we choose the default parameters which are help to best fit the real model without problems of overfitting. And only we need to consider is the kernels and their accuracy. The black points represent the real world oil price of our test data for 90 samples and red line predict value based on our test data. The result comparisons are below and we can see that the Gaussian Process with DotProduct kernel have the highest R-Square and lower MSE.

	MSE	**RMSE**	**R-Square**
RBF&Matern kernels	88.69	9.418	0.8688
RationalQuadratic&RBF	57.59	7.589	0.9148
Constant&DotProduct	126.01	11.225	0.8136
DotProduct	28.53	5.341	0.9578

Table 10. Different kernels performance comparison

Conclusion and Recommendations

In our project. Based on two different time series model like Simple Moving Average and Exponential Moving Average, we got MSE and RMSE which, on average, are higher than machine learning algorithm in producing oil price prediction model. And R-square are lower. Comparing with these three different machine learning methods, Gaussian process regression, super vector regression and random forest, we can see that Gaussian Process with DotProduct have better performance. And the optimal prediction result we figured out is around 0.9578. Since in our model, what we need to do is predict the world crude oil price, therefore, we didn't normalize our y-value(oil price value) and therefore, the mean square error is about 28.53. However, if we normalized the real data, then mean square error will down to around 0.003, which is a trustified value. The comparing dataset is given below.

	MSE	RMSE	R-Square
SMA	124.87	11.174	0.8152
EMA	95.75	9.785	0.8583
GP	28.53	5.341	0.9578
SVR(RBF)	35.96	6.000	0.9468
RF	29.24	5.404	0.9568

Table 11. Comparison on different algorithms

In conclusion, first of all, crude oil price has highly correlation with the features that included, such as the seasonal demand, the previous monthly price index, the world's annual production and consumption, as well as some important world events. Secondly, machine learning models have better performance in crude oil price prediction than the time series model. Finally, Gaussian process regression model with DotProduct kernel gives the best result since it has the highest R-square and lowest MSE and RMSE score.

Admittedly, we still need some more improvement. Firstly, we can add more features, such as the political issues at a specific time, a new policy related to the oil production or shipment and some regulations corresponding to the crude oil market.

Additionally, some new technological inventions or some new alternative energy may also contribute to the lower oil price. And environmental issues may also affect the oil price since it may lead to some environmental pollution. Besides, optimizing the parameters is also a contributor. Since now, we didn't integrate the cross-validation in our gaussian process regression and therefore, we can use this method to help enhancing the test accuracy. Last but not least, we can also consider more data samples like some regional data and weekly data to make our data samples adequate for the deep learning algorithms since we only have the monthly

crude oil price for now. Finally, we can also use some other different models to help us optimize our models such as using some other deep learning methods like neural network.

Reference

[1] M. R. Mahdiani and E. Khamehchi, "A modified neural network model for predicting the crude oil price," *Intellect. Econ.*, vol. 10, no. 2, pp. 71–77, 2017.

[2] M. Gumus and M. S. Kiran, "Crude oil price forecasting using XGBoost," *2nd Int. Conf. Comput. Sci. Eng. UBMK 2017*, pp. 1100–1103, 2017.

[3] A. Khashman and S. Member, "Intelligent Prediction of Crude Oil Price Using Support Vector Machines," pp. 165–169, 2011.

[4] R. A. Ahmed and A. Bin Shabri, "Daily crude oil price forecasting model using arima, generalized autoregressive conditional heteroscedastic and Support Vector Machines," *Am. J. Appl. Sci.*, vol. 11, no. 3, pp. 425–432, 2014.

[5] A. Tebyanian and F. Hedayati, "Intelligent crude oil price forecaster," *Proc. - 2014 13th Int. Conf. Mach. Learn. Appl. ICMLA 2014*, pp. 453–455, 2014.

[6] Y. Zhao, J. Li, and L. Yu, "A deep learning ensemble approach for crude oil price forecasting," *Energy Econ.*, 2017.

[7] W. Xie, L. Yu, S. Xu, and S. Wang, "A New Method for Crude Oil Price Forecasting," *Springer*, vol. 4, pp. 444–451, 2006.

[8] L.Andy, M. Wiener "Classification and Regression by random Forest," *Merck Research Laboratories.*, 2002.

[9] B.Tim, "A conditionally Heteroskedastic Time series model for speculative prices and rates of return" *Review of Economics and Statistics*, 2005.

[10] R.H. Shumway, D.S. Stoffer, "An approach to time series smoothing and forcasting using the EM algorithm," *Journal of Time Series Analysis.*, 1982.

[11] C.Y.Yeh, C-W. Huang, and Shie-J.Lee, "A multiple-kernel support vector regression approach for a stock market price forecasting," *Science Direct.*, 2010.

[12] B.Ash, E.Gerding, and F.M.Groatry "Automated trading with performance weighted random forests and seasonality," *Science Direct.*, 2014.

[13] J. W.-S. Hu, Y.-C. Hu, and R. R.-W. Lin, "Applying Neural Networks to Prices Prediction of Crude Oil Futures," *Math. Probl. Eng.*, 2012.

[14] C.K.Willians, "Prediction with Gaussian processes: from linear regression to linear prediction and beyong," *Technical Report NCRG.*, 1997.

[15] K.-Luckyson, S. Saha, and S-R.Dey, "Predicting the direction of stock market prices using random forest" *Applied Mathematical Finance*, 2016.