Analyze the Growth Rate of Price using Machine Learning

Yuxiao He

University Of Illinois Urbana-Champaign, Urbana, IL, US
Yuxiaoh3@illinois.edu

ABSTRACT
People buy things every day, and the prices of different items change every day for various reasons, such as economic or political issues. This paper will focus on the crude oil price. Crude Oil plays a significant role in the stability of the world economy, and it is also one of the indispensable materials in human modern social life. However, crude oil prices are not always stable, and there are many internal and external factors that determine oil prices. This report will build a Linear Regression Model, and the share price of the oil industry, US dollar index, and the inflation rate will be selected as features. The coefficients of these factors are calculated in order to see how these factors affect the growth rate of oil prices. Each feature selected in this paper is independent of the other. In the real world, these features sometimes affect the price at the same time. Some other features cannot be represented by numbers, so it is very difficult to analyze the growth rate of price. There are many other factors that affect the price of crude oil, so it is difficult to analyze what affects affect more. The linear Regression model might not be the best for analyzing prices. Besides Linear Regression, there are still many models for analyzing prices. Also, the paper did not consider all the problems.

Keywords: *Crude Oil price, Machine Learning, Share Price, Linear Regression*

1. INTRODUCTION

Crude Oil is unprocessed petroleum, and it is a dark brown viscous oily liquid with green fluorescence. Crude Oil is found in geological formations beneath the Earth's surface. It is a mixture of different kinds of liquid hydrocarbons such as alkanes and alkenes. Crude oil is one of the most important sources of energy in the world. Crude oil affects the world economy, and some countries' economic performances depend on the export and production of crude oil.

In Economics, the demand and supply of an item determine the price of that item. Prices of different items vary every day. Sometimes the prices steadily increase, and sometimes the prices suddenly decrease for various reasons. The fluctuations of prices are unpredictable. This paper will focus on the growth rate of crude oil prices.

The crude oil price can be affected by the demand and supply. Besides, there are many other factors affecting the price of oil such as exchange rates, inflation rates, stock market, GDP, etc. The oil price varies over years based on the data in Trading Economics. On this website, it can be seen the price suddenly increased in 2008, when there is an economic crisis. So it is not easy to analyze the variations of price, since some external factors are unpredictable and unexpected.

In recent years, in order to better track and predict the price of crude oil, some researchers analyze the growth rate of price on the basis of establishing consensus forecast.[1] However, the drawbacks of this method are that the consensus comes from different resources, which can distort the original values from other resources.[2]

The current analysis of crude oil prices is mainly to choose a variety of machine learning models to analyze the price growth rate, such as ANN (Artificial Neural Network) and SVM (Support Vector Machine). ANN is a good model for analyzing. ANN stands for Artificial Neural Network, and it can analyze by taking in examples and no programming work with task-specific rules is required.[3] The advantage of ANN is that it can determine the optimal lags in short term, it accurately captures the changing pattern of oil price.[4] The disadvantage of ANN is that the functioning of the network is unexplained, it is difficult to debug.

SVM is another machine learning method used for analyzing. SVM stands for Support Vector Machine. SVM is typically used for classification purposes, but it

can be used to solve regression problems. There are four steps to do SVM-based modeling. The first one is data sampling, then preprocessing the data and dividing the data into in-sample data and out-of-sample data. [5] Next, training the data. Finally, after training the data, the analysis can be done. Many papers use SVM for price prediction. However, when the data sets are relatively large, SVM does not seem to be the best option.

This report analyzes the growth rate of price mainly based on the linear regression models because it is one of the most common and easiest learning models when using it to analyze the price. This paper will try to find the correlations between oil price and share price of oil, the dollar index, and other factors that are responsible for the price of oil. Since Linear Regression can predict prices of crude oil at different dimensions, and it take fewer resources and less time than ANN and SVM, Linear Regression Model seem to be a better way.

2. METHODOLOGY

In linear regression, it computes a "hypothesis" denoted as h using the training set. The hypothesis h takes a feature as input x and we produce the output y which is the oil price.[6] We need to compute the coefficient θ to minimize cost function J(θ). The equation of J(θ) is:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h(x) - y)^2 \quad (1)$$

where m is the number of sample data.

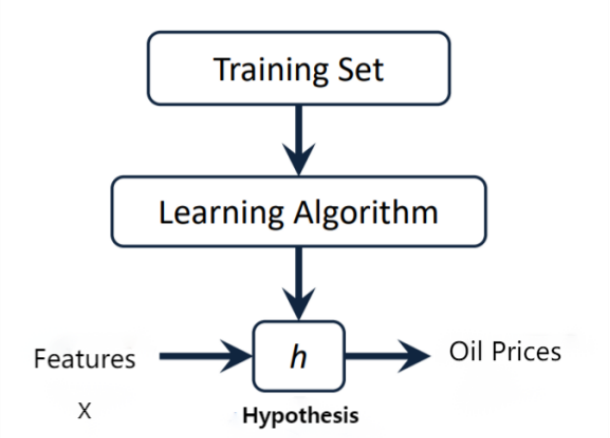

Figure 1: Algorithms of Linear Regression (h means hypothesis)

Linear Regression Model is imported from sklearn. linear_model to calculate the coefficient value θ. Before building the Linear Regression Model, the related data needs to be extracted and processed. The crude oil price is cited from the Kaggle dataset as shown in Figure 2.[7] In one day, there are open price, highest price, lowest price, close price, adjacent close price and the number of shares sold as shown in Figure 2 and Figure 3. As figure 4 shows the data between open price and close price on the same day. The coefficient is 1.0063 which is very close to 1. So there is little difference between the open price and close price which means the price does not fluctuate too much in one day. The coefficient between the high price and low price is also close to 1. So this paper will use the close price for data processing.

Date	Open	High	Low	Close	Adj Close	Volume
22Mar00 – 21Jul20	null 18%	null 18%	null 18%	null 18%	null 18%	null 18%
	29.000000 0%	30.200001 0%	25.900000 0%	26.860001 0%	26.860001 0%	0 0%
	Other (5089) 82%	Other (5090) 82%	Other (5091) 82%	Other (5093) 82%	Other (5093) 82%	Other (5092) 82%
2000-03-22	27.650000	28.250000	27.250000	27.459999	27.459999	92302
2000-03-23	27.650000	27.780001	27.160000	27.309999	27.309999	79373
2000-03-24	27.850000	28.150000	27.549999	27.980000	27.980000	55693

Figure 2: Crude Oil Price (Source: Kaggle dataset)

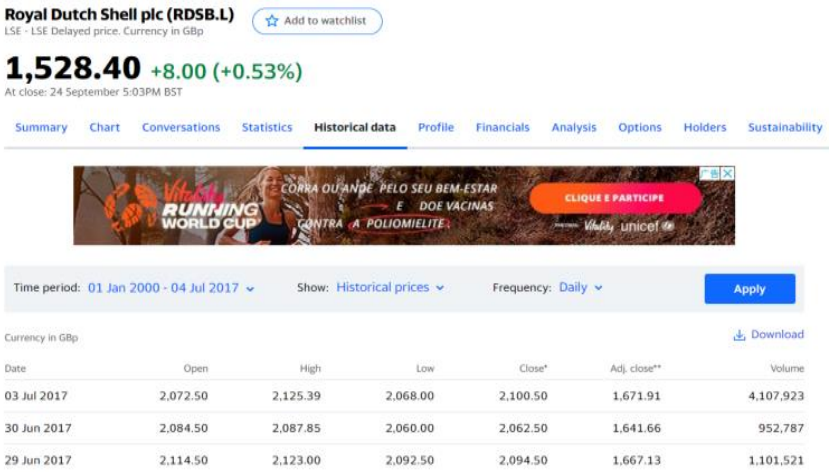

Figure 3: Share price of Oil Industry (Source: Yahoo Finance)

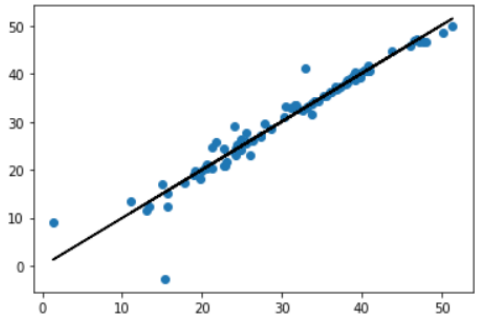

Figure 4: relation between open price and close price

The first feature used as input x is the share price of Royal Dutch Shell. Royal Dutch Shell is one of the biggest oil industries in the world.[8] The reason for using this feature is that the oil industry produces oil, the amount of supply determines the price of oil. The share price comes from Yahoo Finance as shown in Figure 3. [9] There are 100 data in total from the closing price of 2020/02/25 to 2020/07/21. This experiment splits the data into a training dataset and a testing dataset. The ratio of two datasets is eight to two. In Figure 5, the training result is shown. As the share prices of the oil industry increase, the crude oil price will also increase. The coefficient is 0.016, and the mean square error is 90.47.

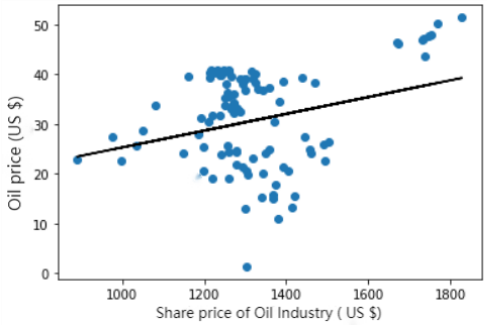

Figure 5: relation between oil price and share price of oil industry

The second feature used as input x is the US Dollar index. US Dollar index can be extracted from investing.com. In Figure 6, the training result is shown. The coefficient is -3.8, and the mean square error is 53.35. As the US dollar index increases, the price of crude oil will decrease.

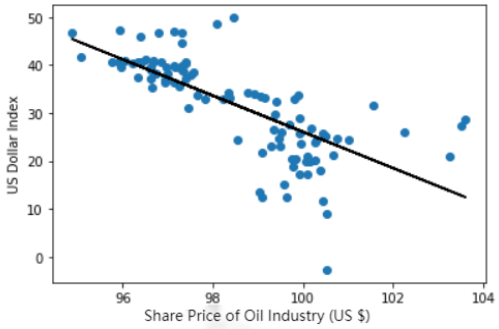

Figure 6: relations between US dollar index and oil price

3. DISCUSSIONS

Two features were chosen in this paper. The coefficients and mean square error of the two features are shown in table 1.

Table 1: Coefficients & means square errors

Feature	Coefficient	Mean Square error
Share price	0.016	90.47
US dollar index	-3.8	53.35

Some features (e.g. Share price of oil industry) are positively affecting the oil price. Some features (e.g. US dollar index) are negatively affecting the oil price. When the absolute value of the computed coefficient is large, it means that the feature affects the oil price more. In table 1, the absolute value of coefficient in US dollar index is larger than the one in share price of oil industry. However, it hard to say US dollar index really has more affections. And there might be relations between US dollar index and share price of oil industry. When combining two features, a polynomial regression model can be built. For example, in table 2, let the sample size m = 100, x_0 is always 1, x_1 is the first feature (share price), x_2 is the second feature (US dollar index), and y is the output price. In polynomial regression, the hypothesis $h = \theta_0 + \theta_1 x_1 + \theta_2 x_2$.[10] The coefficient θ can be calculated in: $(X^T X)^{-1} X^T y$.

Table 2: values of features

x_0	x_1	x_2	y
1	1780.2	98.901	49.9
1	1786	98.938	48.73
…	…	…	…

Also, there are many other features affecting the oil price. More features x_3, x_4 ... can be added when using polynomial regression. However, there are some features that cannot use numbers to represent. For example, it is very difficult to use numbers to describe weather, and the weather may affect the oil price.

There are many factors that can be used as the input features of the Linear Regression model. Therefore, the most affected functions can be found through the application of the linear regression equation, but it takes a lot of time, and the difficulty of obtaining some data will increase the related costs.

There are many factors that affect oil prices, and the instability of oil prices will affect people's daily lives. Therefore, the forecast of oil prices is extremely important. Based on the irregular fluctuations of oil prices, when people buy oil or related products, they will give priority to the timing of purchase. If oil prices are expected to fluctuate many times in a week, consumers can choose to buy at the cheapest price in a week, which can save a lot of money.

4. CONCLUSION

In this paper, US dollar index and share oil price are the features used in the Linear Regression model. Based on the coefficients calculated, US dollar index affects more than share price of oil. Polynomial Regression can be used to combine two or more features.

In conclusion, there are many factors that affect oil price, so it is difficult to analyze what affects affect more. Meanwhile, there might be correlations between each feature. Some factors can indirectly affect the price of oil, and the model might not calculate the desired coefficient. This paper cannot guarantee each feature is independent of the other. In methodology, this paper assumes US dollar index is independent of share price of oil.

The Linear Regression model may not be the best model, there are still many learning models for analyzing the growth rate of price. Here this paper does not consider all the problems, for examples, it is hard to predict the natural disaster. For researches for finding more features, it is better to choose from the perspective of global economics.

ACKNOWLEDGMENT

Firstly, I would like to show my deepest gratitude to my teachers and professors in my university, who have provided me with valuable guidance in every stage of the writing of this thesis. Further, I would like to thank all my friends and parents for their encouragement and support. Without all their enlightening instruction and impressive kindness, I could not have completed my thesis.

REFERENCES

[1] Mikhaylov, A. (2018a), Pricing in oil market and using probit model for analysis of stock market effects. International Journal of Energy Economics and Policy, 8(2), 69-73.

[2] An, Jaehyung. "Oil price predictors: Machine learning approach." 670216917 (2019).

[3] Haykin, S. (1999). Neural Networks: A Comprehensive Foundation, 2nd edition, Prentice Hall, 842

[4] Gupta, Nalini, and Shobhit Nigam. "Crude oil price prediction using artificial neural network." Procedia Computer Science170 (2020): 642-647.

[5] Xie, Wen, et al. "A new method for crude oil price forecasting based on support vector machines."

International conference on computational science. Springer, Berlin, Heidelberg, 2006.

[6] Goyal, Vipor. “MLhandouts”.

[7] Nandan Pandey. “Crude Oil Stock Price.” Kaggle, 22 July 2020, www.kaggle.com/awadhi123/crude-oil-stock-price.

[8] Coupland, Christine, and Andrew D. Brown. "Constructing organizational identities on the web: A case study of Royal Dutch/Shell." Journal of management studies 41.8 (2004): 1325-1347.

[9] “ Share Price of RDSB.L.” Yahoo Finanace, uk.finance.yahoo.com/quote/RDSB.L/history?period1=946684800&period2=1499122800&interval=1d&filter=history&frequency=1d&guccounter=1. Accessed 26 Sept. 2021.

[10] Ostertagová, Eva. "Modelling using polynomial regression." Procedia Engineering 48 (2012): 500-506.