Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230170

Petroleum price prediction based on the linear regression and random forest

Jiaming Bi[1,4,†], Enhui Li[2,†]and Yongxing Luo[3,†]

[1]RCF experimental school, Beijing, China, No.5 North SunpalaceRoad, Chaoyang district, Beijing, China
[2]Harbin No.9 High School, No.2088 ShiboRoad, Songbei district, Harbin, China
[3]Shanghai United International School, No.248 East HongsongRoad, Minhang district, Shanghai, China

[4]BiJiaming@rdfzcygj.cn
†These authors contributed equally.

Abstract. Petroleum price prediction is a challenging task due to the strong volatility in the data. The present study details the development of a predictive model for petroleum prices utilizing two popular algorithms, linear regression and random forest. Based on the analysis of experimental results, it was observed that the model yields an acceptable level of error within the established parameters. Nevertheless, certain limitations were identified, such as the inadequate performance of linear regression in cases where the variation between data points is significant or where the relationship is non-linear. Similarly, random forest algorithms may suffer from reduced accuracy when handling data sets with small sample sizes or low-dimensional data. As a means to address these limitations, the study proposes the application of regularization techniques to mitigate overfitting in linear regression, as well as the handling of null and missing values. Additionally, improvements to random forest performance are proposed, including an increase in the number of trees and the application of hyperparameters to enhance accuracy.

Keywords: petroleum price prediction, machine learning, linear regression, random forest.

1. Introduction

Petroleum, a substance that reacts with other substances to release energy, has been one of the most crucial energies in the world, because crude oil is burned in an engine to generate power that is largely used in war (aircraft, tanks) and also in our daily lives (cars, motorcycle). Economically, the United States and China are the two biggest consumers of crude oil in the world, using 19 million and 16 million barrels everyday respectively, and the largest producer of crude oil is also the United States, where 11,184,870 barrels of petroleum is produced every day. In fact , the prices of crude oil have relied heavily on the supply and demand of it, but in recent years, the war between Russia (the world's second-largest petroleum producer) and Ukraine has affected the demand for petroleum due to sanctions from Europe and the USA and the supply of crude oil is affected by the global pandemic (covid-19) , which resulted in the shortage of labor. Therefore, there have always been fluctuations in petroleum prices, which makes the predictions of prices harder and harder. Nevertheless algorithms in machine learning of artificial intelligence that study the data base can be considered applied in this field to predict the prices of crude oil, which can be served as economic advice and reference to the related companies in order to avoid financial crises.

In the early days, the most traditional methods e.g. financial modeling and the statistics model are widely used to predict the currency, the price of crude oil, gold, valuable luxury and things that does not have a constant price level [1-3]. Although existing models can still provide value and are being utilized, there are notable limitations associated with their use. Specifically, these models commonly lack accuracy in their output and have limited applicability, constraining their utility to certain areas. For example, in the early stages of technology development, researchers may utilize linear regression to create models capable of predicting various outcomes. However, such models often fall short when datasets are not linear or are widely dispersed, resulting in inaccurate results that are challenging to implement. This restricted application of traditional models emphasizes the need for improved methods in order to mitigate the aforementioned challenges.

In recent years, with the development of the world's technology, the artificial intelligence technology has developed in an improving rate and therefore many well-known machine learning algorithms have appeared such as neural network [4, 5], support vector machines [6, 7], decision tree etc. [8, 9]. With the use of these algorithms, people can analyse and process data whose distribution is linear or non-linear. In this paper, a data of current oil prices in a non-linear distribution has been applied to analyse and predict the oil prices in the future. In order to better predict the future oil prices, this paper intends to use advanced machine learning algorithms to accomplish this task.

2. Methodology

2.1. Dataset preparation

In this study, we have chosen to employ a dataset comprising contemporary oil prices, made available by YCHARTS [10]. The oil prices are based on the New York Harbor ultra-low sulfur No.2 diesel spot prices with the unit of USD per gallon collected by the US Energy Information Administration (a statistical agency within the US Department of Energy that regularly collects data concerning energy sources, end uses and energy flows). Spanning as far back as 2006, this dataset has a considerable time frame. Due to the fact that only the recent year of the dataset is required, older data has been eliminated from the dataset that has been employed. After preparing the dataset, the dataset is to be used to make predictions of the oil prices in the future.

2.2. The introduction for machine learning

Machine learning is an artificial intelligence process that involves the use of algorithms to automatically gather and apply data. The systems are expected to analyse the collected data to look for the patterns and use them to make vital decisions for themselves . A machine learning algorithm is the method used by the artificial intelligence system to finish a particular task . Generally speaking , machine learning algorithms predicts values based from the given data . Machine learning can be broadly categorized into four types: supervised learning (e.g. K-nearest neighbour, decision tree, support machine vector, neutral network), unsupervised learning (e.g. K-means clustering, self-organizing map, principal component analysis), semi-supervised learning (e.g. generative models) and reinforcement learning (e.g. Q-learning, multi-armed bandits, Markov decision process). What's more, there are seven major steps in machine learning: collecting data, preparing the data, choosing a model, training the model, evaluating the model, parameter tuning and making predictions. Both collecting and preparing data are important in order to discover the correct patterns using machine learning model. A suitable model must be selected to determine the output values accurately. When training the model chosen , data is passed to the model to find the patterns and make predictions. Then, evaluating the model is done by testing the model. Furthermore, we can try to improve the accuracy of the model which is done by tuning the parameters present in the applied model to make the accuracy closest to its maximum. In the end, the model can be used to make the predictions and finish the task on hand.

2.3. Linear regression

The algorithm used to predict petroleum price is linear regression, which is a machine learning algorithm used to predict numerical variables based on given independent variables. Linear regression which is invented by Sir Francis Galton in 19th-Century models is not only used in the statistic field but also largely applied in the computer science field. In this theory, the y-axis contains the variables that need to be predicted called the dependent variable, and variables used as input to predict the other variables are called the independent variable which is plotted in the x-axis. The model of linear regression is Yi = b0 + b1Xi + ei, in the equation, Yi is the dependent variable, xi is the independent variable, ei is the error items, bo is the intercept and b1 is the slope. People need to plot the graph in advance by using the x and y parameters and they can draw the best-fit line (also called the regression line), then figure out the predicted value based on the line.

It is imperative to use a reliable database in the experiment of linear regression and the database must contain the price of crude oil, which directly affect the price of petroleum and is used in linear regression as the independent variable. Therefore, we set values in the x-axis to the price of crude oil, and we set the values in the y-axis to the price of petroleum. When we are dealing with the database used in this algorithm, we need to read the database in advance by using the read_csv() function, then we can use the function head() to show the assigned number of rows, describe() which is used in calculating numerical values and info() which gives us summary statistics for numerical columns. Last but not least, we use the function drop() to alter the data which we do not want to use. Furthermore, we need to import some packages such as numpy and pandas for basic functions in high level language, For plotting the diagram, we need to import matplotlib. pyplot,seaborn, plotly.graph_objs,plotly,cufflinks. Afterward, we need to import linear regression by using from sklearn.linear_model import LinearRegression.

3. Result and discussion

To make sure the results are fair we decided to use two methods called linear regression and random forest. This can show the difference between the two methods; thus we can compare which one is better.

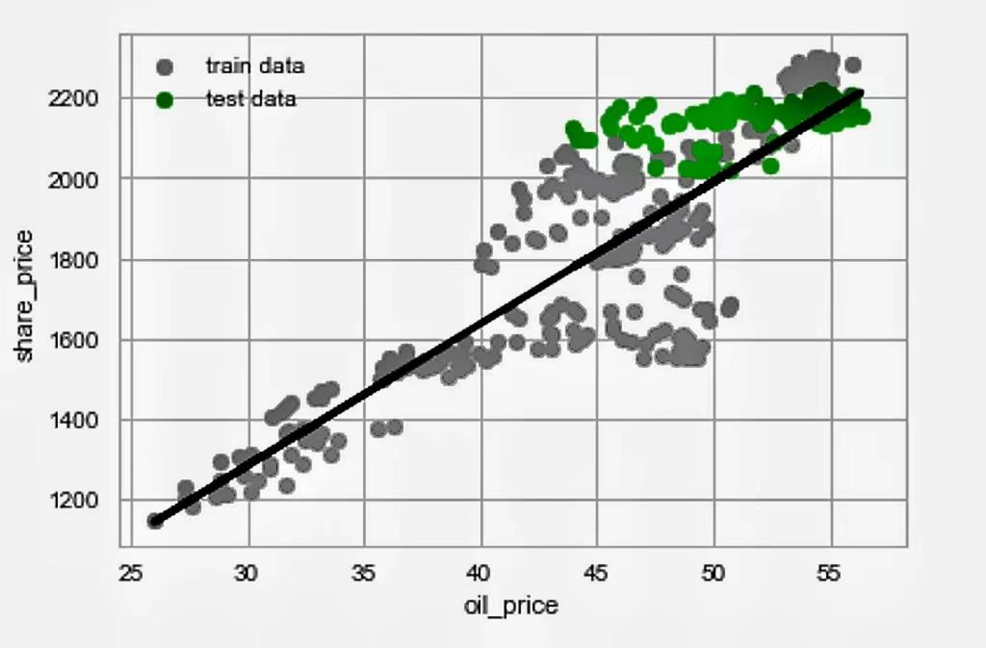

Figure 1. The performance of the linear regression model.

From the Figure 1, it can be observed that by using linear regression is indeed not the best method. The mean squared error is about 23210. The figure shows how the linear regression works by having a line of best fit and using this line to predict the price. As mentioned, if the raw data is scattered then linear regression will have a very low accuracy. By looking at the error and the accuracy it proofed that it does have a very low accuracy.

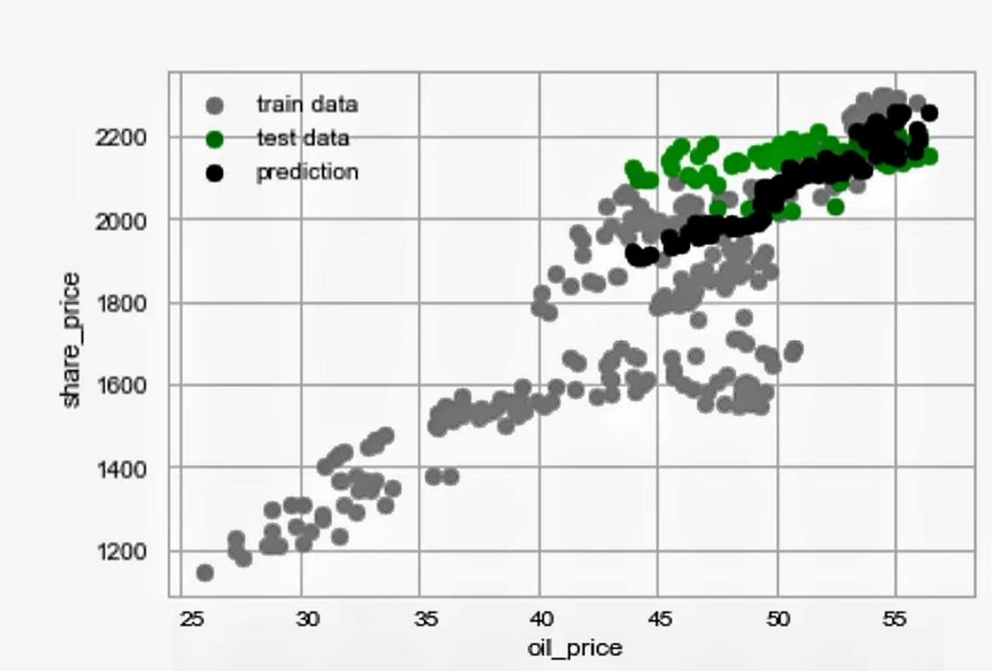

Figure 2. The performance of the random forest model.

Next, we used random forest method. The error decreased enormously, 23210 to 2647 shown in Figure 2. This suggest random forest method is way better than linear regression when the raw data is not lineal. This is because random forest method will take observe set and variable set from the data to build a decision tree and normally it will build more than one decision tree to self-correct which can obtain a higher accuracy.

In summary, it can be inferred that Random Forest model outperforms Linear Regression model in the context of the given dataset, despite a residual error of 2647 which is deemed acceptable when compared to that of the Linear Regression model which is 23210. The superiority of Random Forest model over Linear Regression model is evident through the considerable enhancements made in the transition from the latter to the former. Hence, it can be concluded that Random Forest model is a favorable option for analyzing the given dataset.

4. Conclusion

In conclusion, we finished the model of predicting fuel prices by using linear regression and random forest. By looking at experimental results, we can directly find that there is indeed some error, but it is within the accepted range. However, there are some drawbacks such as the best-fit line in linear regression is not accurate if the difference between two data is too large and it cannot work on the non-linear relationship, in terms of random forest, it is not precise if the data are too small and too low-dimensional data. Therefore, we are planning to use regularization to avoid overfitting and handling the null values and deleting missing values to improve linear regression. Lastly, we are going to improve random forest by increasing the number of trees and using hyperparameters to improve its accuracy.

References

[1] Elshendy M Colladon A F Battistoni E et al. 2018 Using four different online media sources to forecast the crude oil price Journal of Information Science 44(3): 408-421

[2] Reboredo J C 2012 Modelling oil price and exchange rate co-movements Journal of Policy Modeling 34(3) 419-440

[3] Salisu A A Fasanya I O 2013 Modelling oil price volatility with structural breaks Energy policy

52: 554-562

[4] Al-Shayea Q K 2011 Artificial neural networks in medical diagnosis International Journal of Computer Science Issues 8(2): 150-154

[5] Yu Q Yang Y Lin Z et al. 2020 Improved denoising autoencoder for maritime image denoising and semantic segmentation of USV China Communications 17(3): 46-57

[6] Hearst M A Dumais S T Osuna E et al. 1998 Support vector machines IEEE Intelligent Systems and their applications 13(4): 18-28

[7] Noble W S 2006 What is a support vector machine? Nature biotechnology 24(12): 1565-1567

[8] Myles A J Feudale R N Liu Y et al. 2004 An introduction to decision tree modeling Journal of Chemometrics: A Journal of the Chemometrics Society 18(6): 275-285

[9] Quinlan J R 1996 Learning decision tree classifiers ACM Computing Surveys (CSUR) 28(1): 71-72

[10] Ycharts 2023 https://ycharts.com/