Journal of International Technology and Information Management

Volume 27
Issue 4 *Special Issue on ICT for Development (ICT4D)*

Article 1

5-1-2019

Crude Oil Price Prediction with Decision Tree Based Regression Approach

Engu Chen
Fairfield University, engu.chen@student.fairfield.edu

Xin James He
Fairfield University, xhe@fairfield.edu

Crude Oil Price Prediction with Decision Tree Based Regression Approach

Cover Page Footnote

This paper is for the special issue on ICT4D

Crude Oil Price Prediction with Decision Tree Based Regression Approach

Engu Chen
Graduate Student in Business Analytics
Dolan School of Business
Fairfield University
engu.chen@student.fairfield.edu

Xin James He, Ph.D.
Department of Information Systems and Operations Management
Dolan School of Business
Fairfield University
xhe@fairfield.edu

ABSTRACT

Crude oil is an essential commodity for industry and the prediction of its price is crucial for many business entities and government organizations. While there have been quite a few conventional statistical models to forecast oil prices, we find that there is not much research using decision tree models to predict crude oil prices. In this research, we develop decision tree models to forecast crude oil prices. In addition to historical crude oil price time series data, we also use some predictor variables that would potentially affect crude oil prices, including crude oil demand and supply, and monthly GDP and CPI during the period 1992 through 2017 with a total of 312 observations. In this research, we use decision tree models to predict crude oil price. We find that the decision tree models developed in this research are expected to have higher forecasting accuracy than that of such benchmark models as multiple linear regression and time series autoregressive integrated moving average (ARIMA).

KEYWORDS: Crude Oil Price Forecast, Classification Regression Tree, M5P, Random Forest, ARIMA

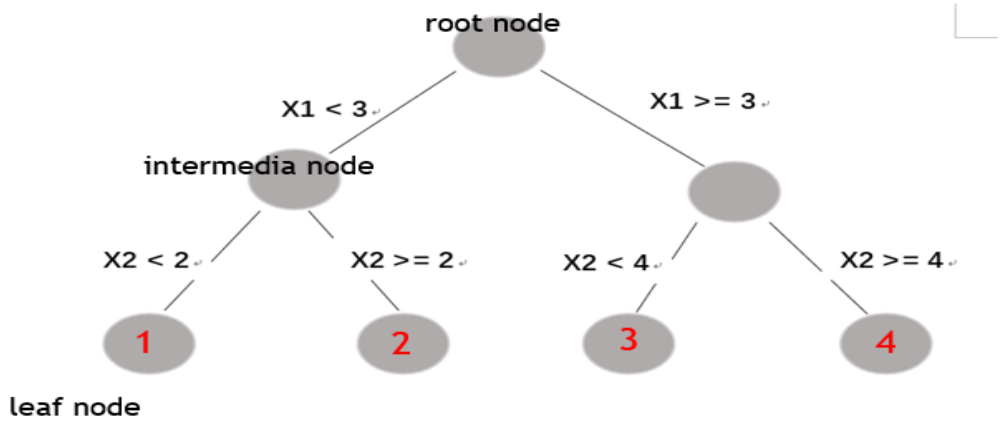
INTRODUCTION

Crude oil is one of the most important commodities for almost every part of the world. The changes in its price have great impact on economies around the world.

Many forecasting methods have been developed to predict the price of crude oil, including conventional econometrics models and machine learning approaches, which can provide more accurate prediction than that of econometrics models but are often difficult to interpret economically. As a popular machine learning approach, decision tree models have great predictive power in some studies. Moreover, unlike other machine learning models that are considered as 'black boxes' due to difficulties to be interpreted economically, decision tree models are interpretable in theory (Loh, 2014). In this research, we develop various decision tree models to compare with such benchmark models as multiple linear regression (MLR) and ARIMA models for forecasting accuracy.

A decision tree is an attractive machine learning method due to its efficiency, robustness, and relatively simple structure (Quinlan, 1986). According to J.R. Quinlan (1992), the prominent advantage of decision tree is that it allows for interpretation easily after making the prediction. The basic decision tree structure can be illustrated in Figure 1.

Figure 1: A Decision Tree Structure



The node on the very top of the tree in Figure 1 is called the root node, which contains the full training dataset where the first split occurs. The nodes at the end of the tree are called leaf nodes, whereas the nodes in between are called intermediate nodes. The root node and the intermediate nodes will split into two subsets based on certain attributes. A decision has to be made whether to split the node into two different nodes or to leave it as a leaf node. This process continues until the tree is fully grown.

ISSN: 1941-6679-On-line Copy

The decision tree models with terminal leaf values categorical are called classification trees, with terminal leaf values numeric are called regression trees, and with terminal values resulting from a linear expression are called model trees (Witten & Frank, 2000).

In this research, we analyze forecasting models by comparing time series models, MLR and ARIMA models, against various decision tree models to improve the prediction accuracy on crude oil prices during the period 1992 through 2017 with a total of 312 observations.

The rest of the paper is organized into sections. Section 2 reviews the literature. Section 3 discusses data collection. Section 4 introduces the methodology. Section 5 presents the analysis and results. Finally, Section 6 provides the concluding remarks.

LITERATURE REVIEW

Decision Tree Algorithms

Breiman, Friedman, Olshen, and Stone (1984) proposed Classification and Regression Tree (CART) in their book *Classification and Regression Tree*. They defined the regression model as Tree Structured Regression to differentiate it from other regression methods, where the training set is partitioned by a sequence of binary splits into terminal nodes. In each terminal node, a numerical value will be generated as the predicted value at each leaf node. Consequently, they came up with three specific rules to determine a regression tree model.

The first rule is how to select a split at the root node and intermediate nodes. In CART models the criterion to split at a node is to decrease the variance of that node the most. Suppose we have a training dataset T in a node containing n observations with continuous predictor variables x_1, x_2, …, x_i, and a continuous target variable y. We first try to split on x_1 into two subsets, denoting the subsets on the left as T_L and on the right as T_R. The number of observations in each subset as a and b, respectively. Let V(T) be the variance of the target variable in the original dataset, and $V(T_L)$ and $V(T_R)$ be the corresponding variances of the two subsets. Hence, the variance reduction can be computed by:

$$\Delta V(T) = V(T) - [\frac{a}{n} V(T_L) + \frac{b}{n} V(T_R)] \quad (1)$$

By the same token, Eq.(1) can be used for the variance reduction on all the predictor variables x_i so long as the split criterion that generates the highest variance reduction is used for each of the nodes. In general, the unique values in each attribute will be sorted and treated as discrete values with the average of two adjacent unique values as a potential split point for continuous predictor variables. Hence there will be k-1 possible splitting points in an attribute with k unique values.

The second rule is how to determine when a node is terminal. The training set can split successively until a) the length of every branch reaches the predetermined threshold or b) the number of observations in every node reaches the predetermined threshold. When all the nodes are terminal, the model prunes the tree before it is finalized by balancing model complexity and prediction accuracy.

The third rule is how to assign a value to a terminal node. To minimize the prediction error, the predicted value at a terminal node in CART models is the average of the target variable values resulting from the observations at that node.

In order to improve the prediction accuracy, Breiman et al (1984) suggested that the CART models can be enhanced by replacing the average of the target variable values with a linear regression, which leads to M5 models (Quinlan, 1992). The rules used in CART models can be applied to M5 models if the standard deviation is used, instead of the variance as in the case of the CART models. Let sd(T) be the standard deviation of the target variable in the original dataset, and sd(T_L) and sd(T_R) be the corresponding standard deviations of the two subsets. Consequently, Eq. (1) becomes Eq. (2)

$$\Delta sd(T) = sd(T) - [\frac{a}{n} sd(T_L) + \frac{b}{n} sd(T_R)] \quad (2)$$

The split that generates the highest standard deviation reduction in M5 models will be used as the split criterion of the node.

According to Quinlan (1992), there are 5 major improvements of the M5 models over the previous decision tree models such as CART.

1. To deal with potential underestimate of the residual in terms of mean absolute error in the training set, M5 models multiply the residual of linear models by a factor of $(n+v)/(n-v)$, where n is the number of observations in the training set and v is the number of parameters in the linear model at each leaf node.
2. A multivariable linear model is constructed to restrict the attributes that are referenced by tests or linear models somewhere in the subtree at this node

ISSN: 1941-6679-On-line Copy

to ensure the linear model from the node has the same accuracy as that from the subtree.

3. After each linear model is obtained, M5 will simplify these linear models by removing some no-contributing variables by a greedy search method.

4. Each non-leaf node is examined starting near the bottom to see whether it is possible to replace a subtree with a linear model so that this node may be pruned to a leaf with a better estimated error.

5. To compensate for the discontinuous nature of the linear models at adjacent nodes, a smoothing process is conducted to improve the prediction accuracy of the tree-based model. For each subtree, beginning from the bottom, denoting the node on the top as S and nodes below as S_i, PV(S) as the predicted value at S, n_i as the number of observations in node S_i, and M(S) as the predicted value given by the linear model, the predicted value of all observations in S is computed by:

$$PV(S) = \frac{n_i * PV(\mathrm{Si}) + k * M(S)}{n_i + k} \quad (3)$$

where k is a smoothing constant with a default value of 15.

Quinlan (1992) concluded that a model tree can learn effectively from large data like a regression tree. In addition, a model tree has advantages over a regression tree in terms of forecasting accuracy because a regression tree never gives predicted value outside the range of training cases while a model tree can via extrapolation.

Wang and Witten (1997) developed certain improved M5 models and named them M5 Prime (M5P) models with the following characteristics:

1. In the M5P model, a leaf node with fewer than 2 training examples will automatically trigger a pruning and consequently will not exist in the final model. Also, the M5P will not split a training set if it contains 3 or fewer examples in the initial split whereas this threshold in M5 is 2. Neither does the M5P split the subset if the standard deviation of the subset is less than 5% of the original training set.

2. Unlike in M5 models, non-contributing predictor variables will not be removed from the M5P models since the authors believe removing them makes no significant difference.

3. The M5P compares the accuracy between the linear model at the node and the weighted linear sum from the subtrees to decide the necessity of pruning.

Breiman (2001) found that the random forest (R_Forest) is a model that ensembles all the decision tree models to come up with the strongest model as its output. With a large number (usually hundreds) of randomly generated trees, the accuracy of the random forest is very likely to be better than any individual regression tree model (Breiman 2001). The random forest regressor will generate an ensemble of trees and then vote for the most popular class.

To maximize the predictive power of the random forest, the model must balance between the strength of each tree and the correlation of the trees. With more features and observations selected, the accuracy of each regression tree model will be stronger, but the correlation between these trees will also be higher. Stronger individual trees will lead to better overall accuracy of the random forest algorithm, while higher correlation between them will lead to higher model complexity. Moreover, a high correlation between individual trees might also reduce the accuracy of the random forest model.

Research studies with the goal of finding factors influence crude oil prices can be roughly categorized into three categories. Studies focus on traditional time series models such as ARIMAX (Elshendy, Colladon, Battistoni, and Gloor, 2017), an extension of the ARIMA models, on machine learning models such as Artificial Neural Network (ANN) models (Abdullah and Zeng, 2010), and on both time series and machine learning models such as Supporting Vector Machine (SVM) (Xie et al, 2006; He, 2018). Often, ARIMA is used as a baseline model for comparison with respect to crude oil price forecasting (He, 2018).

Nwulu (2017) conducted a research on crude oil price prediction using decision tree based models. The author used weekly crude oil option prices from January 2, 1986 to October 21, 2013, to train 5 different models, including M5P, Random Forest and some other decision Tree models, of which M5P model was claimed to have the best performance. While option prices can be derived from crude oil prices, these results cannot be compared directly with monthly crude oil price forecasting in this research in terms of model accuracy.

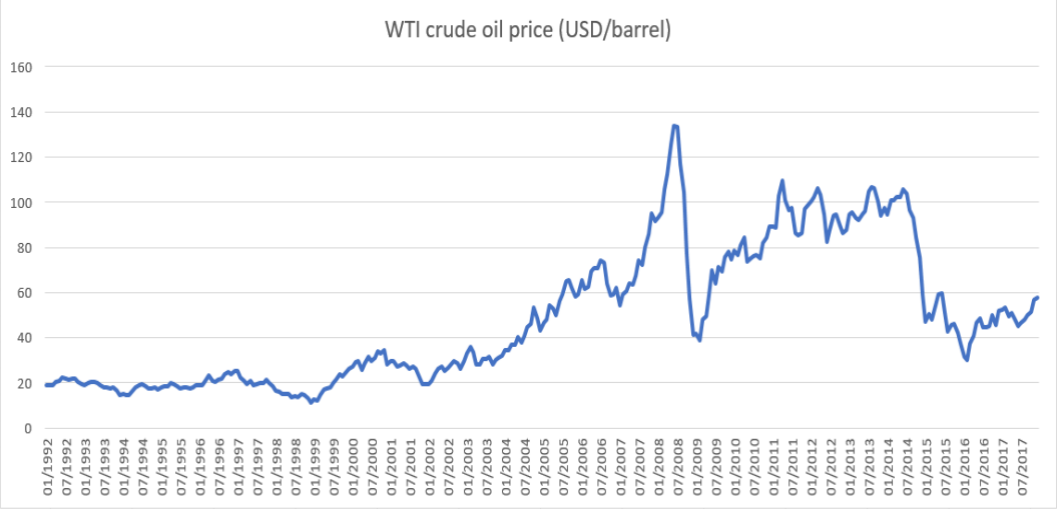
ISSN: 1941-6679-On-line Copy

DATA COLLECTION

Spanning from 1992 January to 2017 December, we collected 312 observations of monthly data with eight predictor variables on crude oil prices as the target variable. The time series of the U.S. Energy Information Administration on West Texas Intermediate (WTI) spot prices was used as the target variable, along with four of the eight predictor variables: World Crude Oil Production (WorldProduction), oil consumption of the Organization for Economic Co-operation and Development countries (OECDConsumption), U.S. oil production (USProduction), and U.S. oil consumption (USConsumption). Two of the predictor variables were from the U.S. Bureau of Labor Statistics: U.S. Gross Domestic Production (GDP) and U.S. Consumer Price Index (CPI), and the other two predictor variables were from the U.S. Federal Reserve: U.S. Dollar Exchange Index (DollarIndex) and U.S. Federal Reserve Interest Rate (InterestRate).

To capture the autocorrelation between the monthly WTI crude oil spot price and its values of the previous months, we also included the time series (Time) and lagged crude oil prices L1, L2, L3, L6 and L12 as predictor variables, where L1 stands for one month lagged oil price and L2 for two months lagged, and so on. Figure 2 shows the monthly time series plot of the crude oil prices during the period 1992 through 2017 with a total of 312 observations.

Figure 2: Time Series Plot of the Crude Oil Prices 1992 – 2017



METHODOLOGY

With the target variable and all the predictor variables being continuous, we now deploy CART regression tree, Random Forest Regressor, and M5P model to compare their model performance and accuracy. MLR and ARIMA model are used as benchmarks for comparison. We use R to run all the models in this research.

Specifically, we use R package rpart to model CART and R package RWeka to analyze the M5P models, which fits linear regression models in its leaf nodes and is expected to outperform the CART models. In addition, we use R package randomForest to generate the Random Forest Regressor, which usually will have higher forecasting accuracy than any of the single decision tree models due to its ensemble nature.

The forecasting model accuracy is often measured by the level of prediction error: the lower the prediction error, the better the forecasting accuracy. To compare the performance of each of the models, we used Root Mean Squares Error (RMSE) and Mean Absolute Error (MAE), along with R^2, which measures the explanatory power of the model towards the data. The lower the value of RMSE and MAE, the better the model accuracy, and the higher the R^2, the better explanatory power the model.

$$\text{RMSE} = \sqrt{\frac{\sum_1^n (A-T)^2}{n}}$$

$$\text{MAE} = \frac{\sum_1^n (A-T)}{N}$$

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

ANALYSIS AND RESULTS

Table 1 compares 3 decision tree models against the benchmark MLR and ARIMA modes.

ISSN: 1941-6679-On-line Copy

Table 1: Forecasting Accuracy of Various Models

	M5P	R_Forest	CART	MLR	ARIMA
R^2	0.9984	0.9966	0.9688	0.9830	0.9799
MAE	2.5672	1.2481	4.8093	2.8439	2.9444
RMSE	3.6410	1.8613	6.9553	3.9452	4.2886

As shown in Table 1, M5P, R_Forest, and CART are decision tree models, whereas MLR and ARIMA are time series models. It is seen in Table 1 that the original Classification and Regression Tree (CART) model performs the worst among all five models in this research, with $R^2 = 0.9688$, MAE = 4.8093, and RMSE = 6.9553, which does worse than the benchmarking MLR and ARIMA. Since M5P is an improved CART model, it is no surprise that M5P outperforms CART, and even slightly better than the benchmark models of MLR and ARIMA, with $R^2 = 0.9984$, MAE = 2.5672, and RMSE = 3.6410. It is seen from Table 1 that R_Forest is the best model of all the five models in this study with MAE = 1.2481 and RMSE = 1.8613 although its $R^2 = 0.9966$ is not as high as $R^2 = 0.9984$ of M5P. However, the difference of 0.9986 - 0.9966 = 0.0018 in R^2 is negligible in terms of the explanatory power when both R^2 are greater than 0.99. The MAE = 1.2481 and RMSE = 1.8613 of the R_forest model are less than a half of these of the benchmarking models MLR (column 5) and ARIMA (column 6) in Table 1 above respectively, indicating that the R_Forest model has the best forecasting accuracy.

Figure 3 illustrates how the CART splits the data into subsets. The data set at the root node splits into two subsets at the L1 value of 57.8. The subset with L1 less than 57.8 on the left, for example, further splits into two subsets again at the L1 value of 34.6, which leads to a target value prediction of 47.4 with a 17.9% of the data fitted in this leaf node when L1 is greater than or equal to 34.6. This model generates only five numeric target values of 18.6, 28.9, 47.4, 69.6 and 98.6, with corresponding probabilities of 30.1%, 17.0%, 17.9%, 16.7%, and 18.3%, respectively. In other words, the small number of leaf nodes leads to a limited number of predicted target values, which makes the CART model the worst among the five forecasting models in this research.

Figure 3: The Mechanism of the CART Model

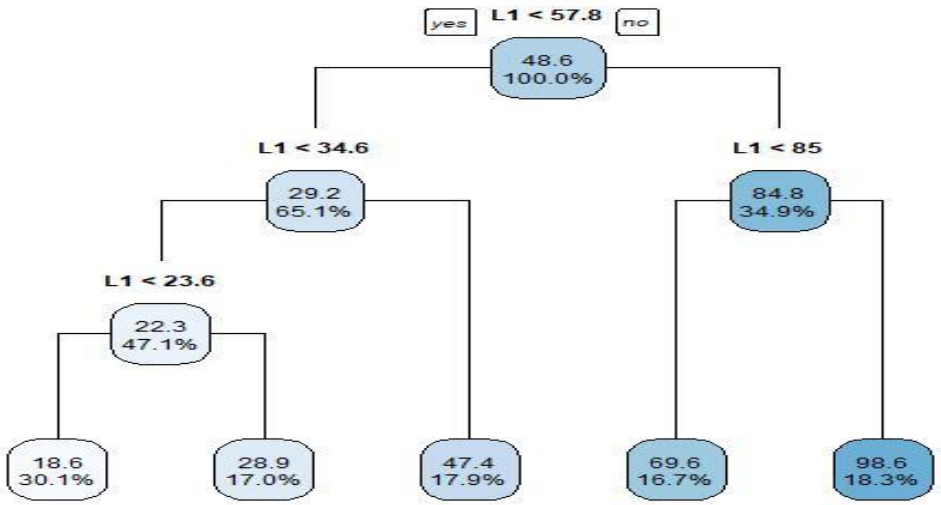
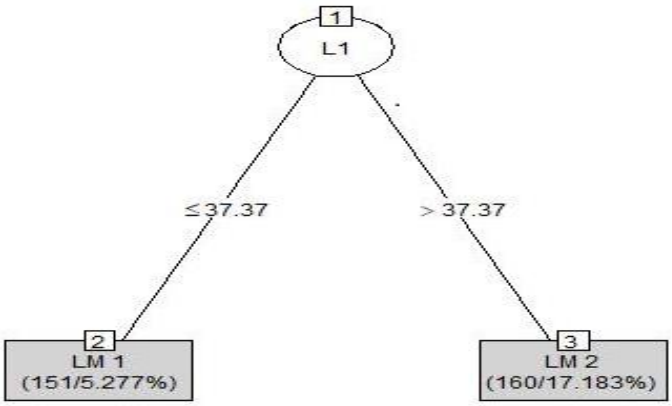


Figure 4 demonstrates the mechanism of the M5P model, where it splits the training dataset, at the L1 value of 37.37, into two linear models, LM1 and LM2. For L1 less than or equal to 37.37 with 151 observations, LM1 has a root relative squared error of 5.277%, which is the normalized average of the actual values from this simple predictor; for L1 greater than 37.37 with 160 observations, LM2 has a root relative squared error of 17.183%.

ISSN: 1941-6679-On-line Copy

Figure 4: The Mechanism of the M5P Model



In general, different split criteria will lead to different subsequent linear models, which in turn result in different predicted target values. Specifically, the resulting LM1 has the following model parameters:

WTI = -0.0164Time + 0.0006USConsumption - 0.0001GDP + 0.0713CPI –
0.0557DollarIndex + 1.1155L1- 0.3183L2 + 0.1265L3 - 11.9294

And LM2 has the following model parameters:

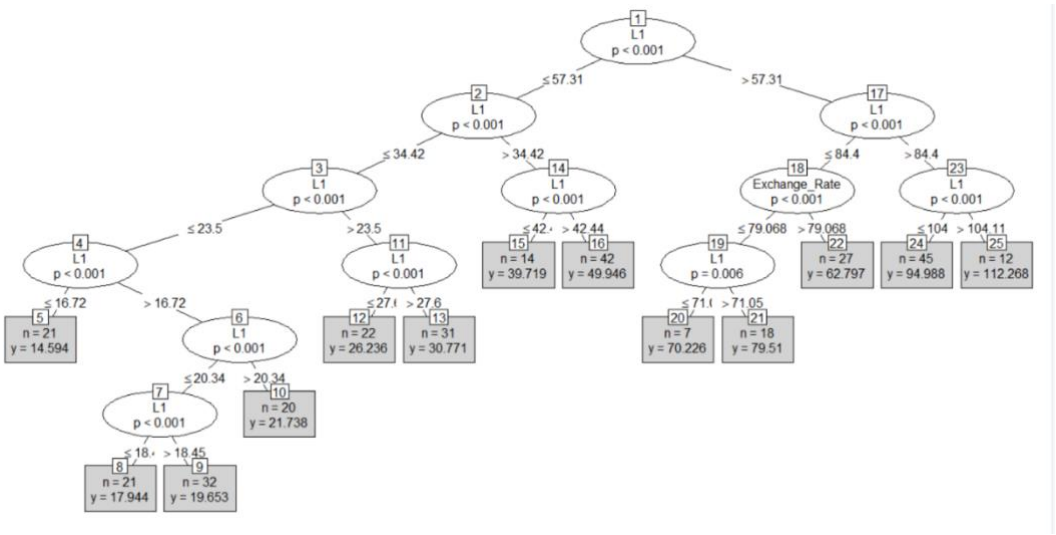
WTI = -0.3181Time + 0.0011USProduction + 0.0001USConsumption +
0.025GDP + 0.636CPI - 0.8396DollarIndex + 1.0038L1 - 0.1472L2 –
0.1653L3 - 23.5343

Thus, the M5P models can capture the linear relationships between the target variable and the predictor variables with linear models. In addition, since the predicted target values in M5P can be any values resulting from the linear models (such as LM1 and LM2 in Figure 4), the predicted target values will have much wider range compared with that of the CART model in Figure 3, which makes the M5P more accurate for the testing set that may not be in the range of the training set.

ISSN: 1941-6679-On-line Copy

Figure 5 presents a sample random forest model since it is impossible to show as many as 100 trees as a complete random forest model. The prediction process of the random forest model can be summarized below.

Figure 5: The Aggregated Rule of a Random Forest Model



Like a model tree such as M5P, the predicted target values are the results of the linear models from the leaf nodes. Unlike the regression tree, the random forest model would generate not just one, but many regression trees known as ensemble nature, and vote for the best as the final forecasting model, which provides potential structural superiority over the rest of the decision tree models in terms of forecasting accuracy.

The random forest model seen in Figure 5 has a total of 13 leaf nodes, a lot more leaf nodes than in CART and M5P. The root node N1 splits at the L1 value of 57.31 into node N2 on the left for L1 less than or equal to 57.31 and note N17 on the right for L1 greater than 57.31. Then node N17 splits further at the L1 value of 84.4 into note N18 on the left for L1 less than or equal to 84.4 and note N23 on the right for L1 greater than 84.4. By the same token, node N23 splits at the L1 value of 104.11 into two leaf nodes with predicted target values of 94.988 on the left and 112.268 on the right.

In addition, Random Forest model can generate variable importance. For the model in Figure 5, the variable importance is as follows:

ISSN: 1941-6679-On-line Copy

Table 2: Variable Importance of Random Forest

	%IncMSE
Time	10.477557
OECDConsumption	6.394415
WorldProduction	7.814262
USProduce	7.491716
USConsumption	8.215248
GDP	10.804185
CPI	9.173369
DollarIndex	15.394203
InterestRate	9.243012
L1	22.331074
L2	14.099587
L3	11.042187
L6	8.409004
L12	6.260562

The variable importance provides a list of the %IncMSE, which measures the percentage increase in mean squared error of the predicted value as a result of the variable being permutated. The higher the number, the more important. It is seen from Table 2 that the one month lagged crude oil WTI prices of L1 has the greatest impact on the value of the target variable with a rate of 22.331074, whereas the U.S Dollar exchange rate has the second greatest impact with a rate of 15.394203. It is no surprise to see L12, the 12 months lagged WTI prices, has the least impact with a rate of 6.260562, but it is counter-intuitive to find that OECD Consumption has the second least impact with a rate of 6.394415.

CONCLUDING REMARKS

In this research, we compared three decision tree models, Classification and Regression Tree (CART), M5 Prime (M5P), and Random Forest, with two benchmarking multiple linear regression (MLR) and time series autoregressive integrated moving average (ARIMA) models on West Texas Intermediate (WTI) spot oil prices as the target variable during the period 1992 through 2017, with 8 predictor variables. Forecasting accuracy is measured in terms of r squares (R^2), root mean squared error (RMSE), and mean absolute error (MAE).

ISSN: 1941-6679-On-line Copy

We found that the Random Forest model is the most superior among all five models studied in this research in terms of forecasting accuracy. Based on the variable importance rankings provided by the Random Forest model, we found that the one month lagged WTI, L1, has the greatest impact on WTI, and the US dollar exchange rate index has the second greatest impact on WTI, indicating WTI spot prices are influenced not only by its previous month WTI spot price but also by the US dollar exchange rate index. Moreover, we found that the 12 months lagged WTI, L12, has the least impact on WTI spot prices, indicating a non-seasonal pattern of the WTI spot prices. However, we are unable to explain the fact that the OECD Consumption has the second least impact on the WTI spot prices, which is counter-intuitive and deserves further investigation in the future.

It is worth noting that the original classification and regression tree (CART) should not be used for predicting continuous target variables due to its lowest forecasting accuracy among all five models in this research. In fact, the CART model in our research performs much worse than the two benchmarking models, MLR and ARIMA. As to the M5P model, it is somewhere in the middle, slightly better than the MLR and ARIMA models but not as good as the Random Forest model.

Future research may try to explain why the OECD Consumption of oil products did not impact the WTI spot prices as much as we all intuitively think it should be. Will longer time series with more observations or more predictor variables help?

REFERENCES

Abdullah S.N. and X. Zeng (2010). Machine learning approach for crude oil price prediction with Artificial Neural Networks-Quantitative (ANN-Q) model. *The 2010 International Joint Conference on Neural Networks* (IJCNN), Barcelona, 1-8.

Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen (1984). *Classification and Regression Trees*. Belmont, Calif.: Wadsworth International Group.

Breiman, L. (2001). Random Forests. Machine Learning, October 2001 (45), 5-32.

Elshendy, M., Colladon, A. F., Battistoni, E., & Gloor, P. A. (2017). Using four different online media sources to forecast the crude oil price. Journal of Information Science, 44 (3), 408-421.

ISSN: 1941-6679-On-line Copy

He, Xin James (2018). Crude Oil Price Forecasting: Time Series vs. SVR Models. Journal of International Technology and Information Management, 27 (2), 25-43.

Ivan H. Witten & Eibe Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA

Loh, W. (2014). Fifty Years of Classification and Regression Trees. International Statistical Review, 82 (3), 329-348.

Nwulu, N. I. (2017). A decision trees approach to oil price prediction. *2017 International Artificial Intelligence and Data Processing Symposium (IDA*P), Malatya, 2017, 1-5.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*,1 (1), 81-106.

Quinlan, J.R. (1992). Learning with Continuous Classes. *Proceedings of Australian Joint Conference on Artificial Intelligence*, Hobart, November 16-18, 1992, 343-348.

Wang, Yong & Witten, Ian. (1997). Inducing Model Trees for Continuous Classes. Proceedings of the Ninth European Conference on Machine Learning, 128 – 137.

Xie, W., L. Yu, S. Xu, and S. Wang (2006). A New Method for Crude Oil Price Forecasting Based on Support Vector Machines, International Conference on Computational Science (Part IV), 444-451.