Impact of News Sentiment on Stock Market

Yuchen Cao[a,b], Chang Li[a,b], Guanqian Wang[a,b], Feng Yu[a,b]

[a] *Department of Mathematics, University of Central Florida*
[b] *Team Name: UCF MATH*

Abstract

In this report, we investigate the impact of news sentiment on the stock market. Specifically, we use the stock price of energy companies like Exxon as the target variable, and news sentiment score, major market indicators like Dow Jones, crude oil commodity price and some other features as the independent variables. Using the traditional linear regression and step-wise model selection, we find that the crude oil commodity price and major market indicator indexes have strong linear correlation with the stock price; however, the variable news sentiment is not considered useful in such a model setting. We also find that residuals of the linear model still follow some interesting patterns. Thus, we decide to move on to using Neural Network to investigate the relationship between the stock price and the news sentiment.

1. EDA and Linear Model Approach

1.1. Introduction

At this section, we perform some basic exploratory data analysis on the given sample data set and some other data set obtained following the manner of sample data set. Also, we implement some fundamental statistical methods like linear regression to make the inferences about the relationship between the stock prices, news sentiment scores and the other features associated with the stock prices. The main tool for this section is R.

We do not expect to make a model with strong ability to predict stock price at this section. In a matured market, there should be no opportunity for arbitrage. If there's a traditional statistical model can successfully predict the stock prices, an arbitrage opportunity will be created, which contradicts the no arbitrage assumption. We consider the no arbitrage assumption is reasonable because the large energy companies are usually watched by many traders and the market should be matured.

1.2. Exploratory Analysis of Sample Data

1.2.1. Exxon News Sentiment Score

In this section, we demonstrate some basic visualizations of the sample data and provided some interpretation. From the Figure 1's left picture, one can see that the sentiment score of news sentiment for Exxon is roughly following a normal distribution. From the left picture of 1, one may also see that the news sentiment score is evenly distributed around the line $x = -0.2$, besides a few outliers. This means one may assume that the news sentiment score is following a normal distribution for Exxon.

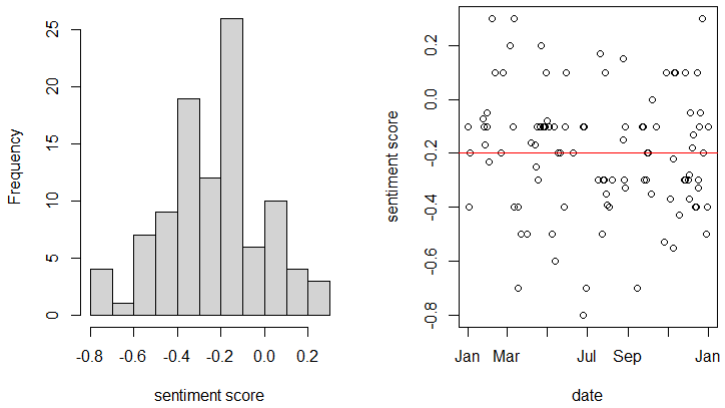


Figure 1: The left figure is the histogram of news sentiment score for the Exxon during year 2020; the right hand side figure is the news sentiment score versus Date of 2020.

1.2.2. Exxon Stock Prices and Crude Oil Commodity of 2020

In this section, the Exxon stock prices, crude oil commodity settle prices from NYMEX and Ice Brent are investigated. From the Figure 2, one can see that Exxon closing price, Ice Brent and NYMEX settle prices are following a very similar trend, except that NYMEX has a sudden negative settle price. Note that the negative settle price from NYMEX is not a wrong observation. NYMEX crude oil settle price did drop to negative during 2020. The Figure 3 is the histograms of three prices, and one can see that they are rather skewed compared with normal distribution; however, note that the number of observation is beyond 200 for all three prices. By central limit theorem, one may still assume they follow normal distribution after some proper normalization.

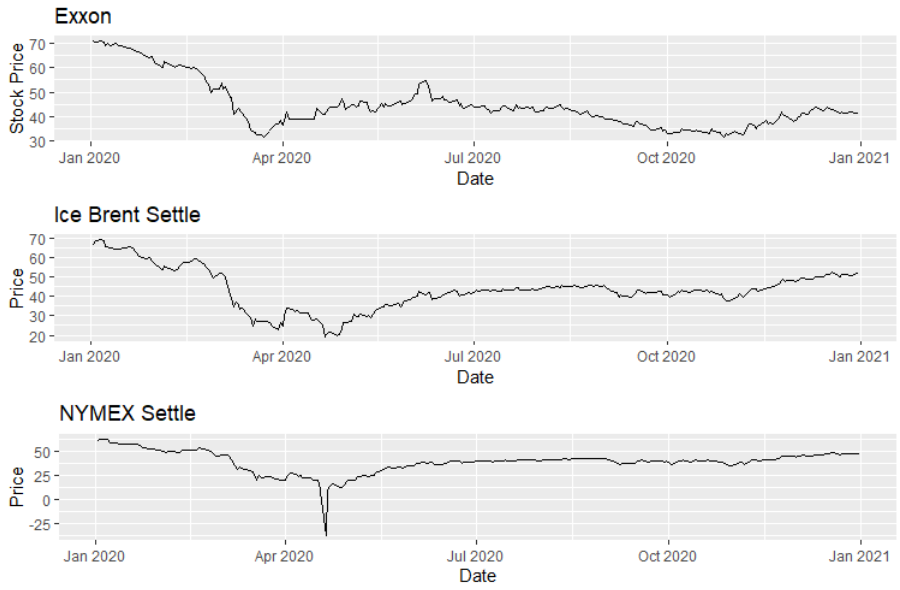


Figure 2: The price of Exxon closing price, Ice Brent settle prices and NYMEX settle price vs. Date of the year 2020.

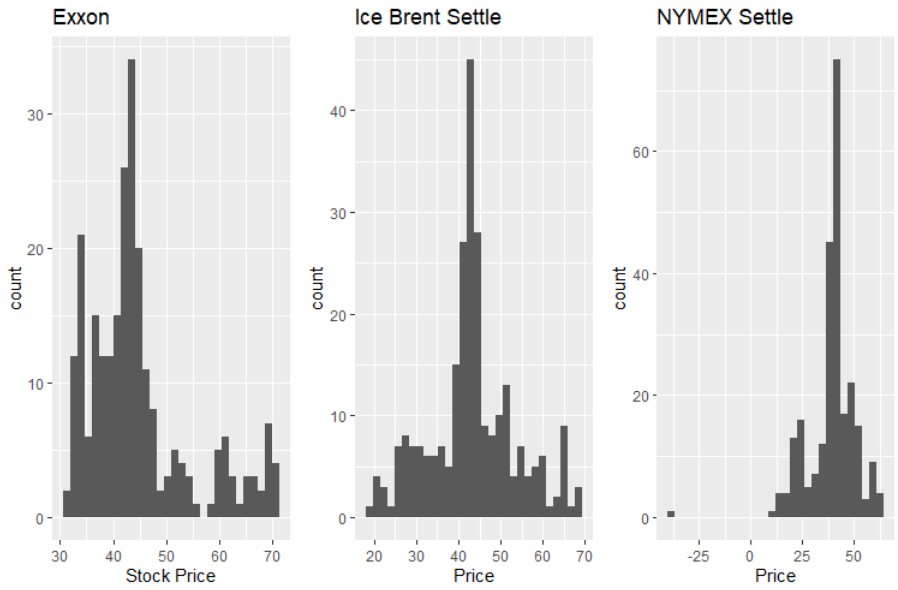


Figure 3: The histograms of Exxon stock closing prices, Ice Brent settle prices and NYMEX crude oil settle price.

1.2.3. Dow Jones and SP 500

In this section, the report investigate the distributions and patterns of several Dow Jones Indexes and SP 500. From the Figure 4, one can see that the trend of price changes of 4 different Dow Jones' indexes are following the same pattern. The Figure 5 shows the histograms of 4 different indexes. One can see that they are not exactly follow the normal distribution, but with some proper normalization, one may still consider they are normal distributed. The similar situation can be observed in Figure 6.

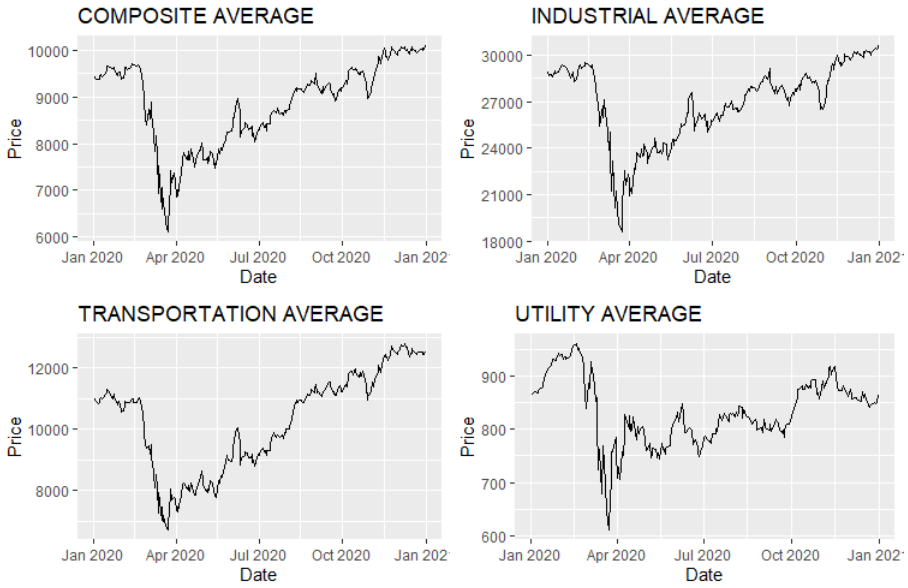


Figure 4: The price of several Dow Jones Indexes vs. Date

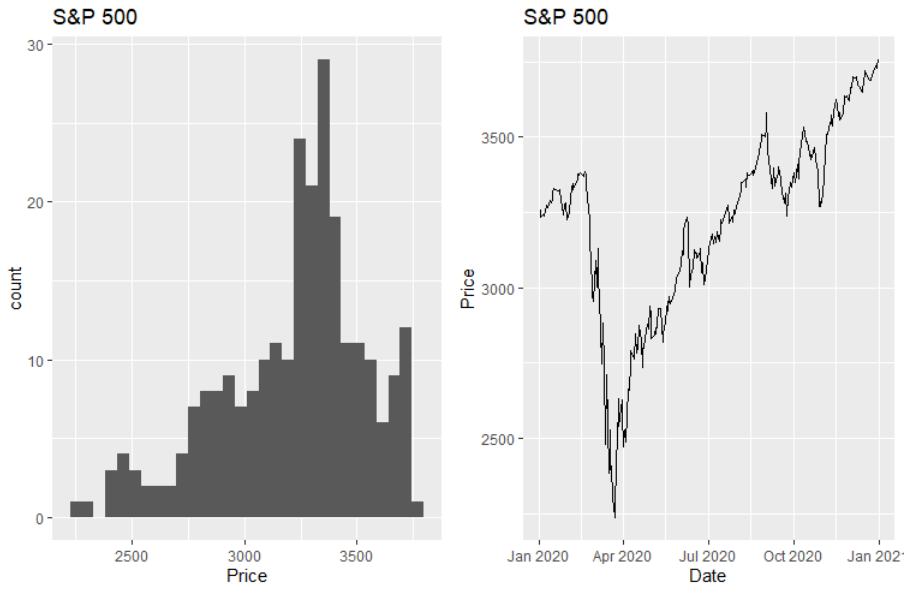


Figure 6: The left plot is the histogram of SP 500 prices of 2020; the right plot is the price vs. date of SP 500 of 2020.

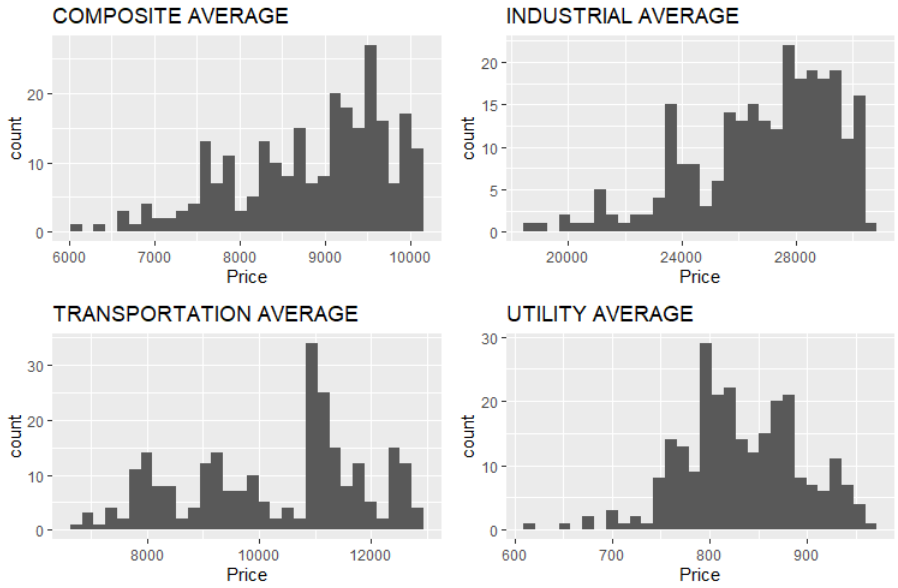
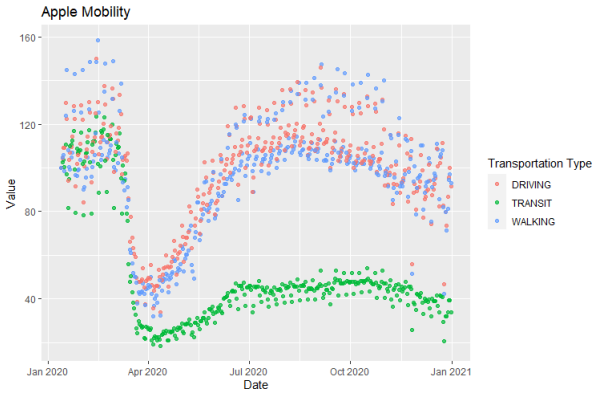


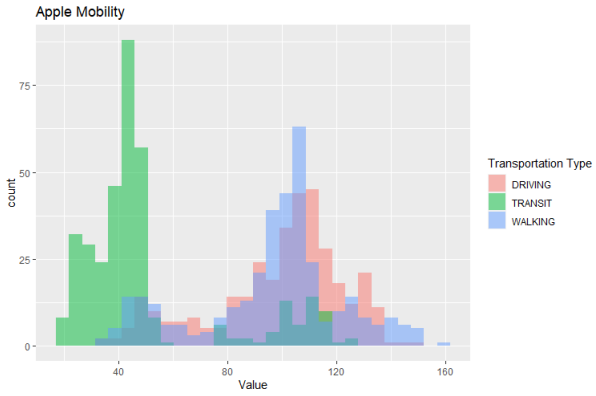
Figure 5: The price of several Dow Jones Indexes' histogram

1.2.4. Apple Mobility

This section provides some visualizations of the Apple mobility indexes. This report skips the Google Mobility scores. The Google Mobility is rather complicated and we have find the Google Mobility not particular useful as a predictor to making inference of stock price. The Figure 7 shows the histograms and trend of different Apple Mobility indexes. The normal assumption for this set of data is also applicable with the same reason as in previous sections.



(a) Apple Mobility vs. Date



(b) Apple Mobility histogram

Figure 7: Apple Mobility

1.2.5. Pairwise Correlation

The report provides a pairwise correlation of all the features investigated above. From Figure 8, one can see that the the news sentiment is somewhat correlated with the stock price, Dow Jones Average Utility and Apple Transit Index. One may also see that Dow Jones indexes are strongly correlated with each other and SP 500. One may also observe that the stock price of Exxon is strongly associated with the Ice Brent/NYMEX Crude Oil settle price. For this graph, the measure of the correlation is the Pearson's correlation. Since it is a classical statistic, this report does not show details of how it is computed.

1.3. Methodology and Assumptions

1.3.1. Normalization

We decide to normalize every feature mentioned from the previous section. The normalization method is:

$$\mu = \sum_{i=1}^{n} \frac{x_i}{n} \tag{1.1}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \tag{1.2}$$

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1.3}$$

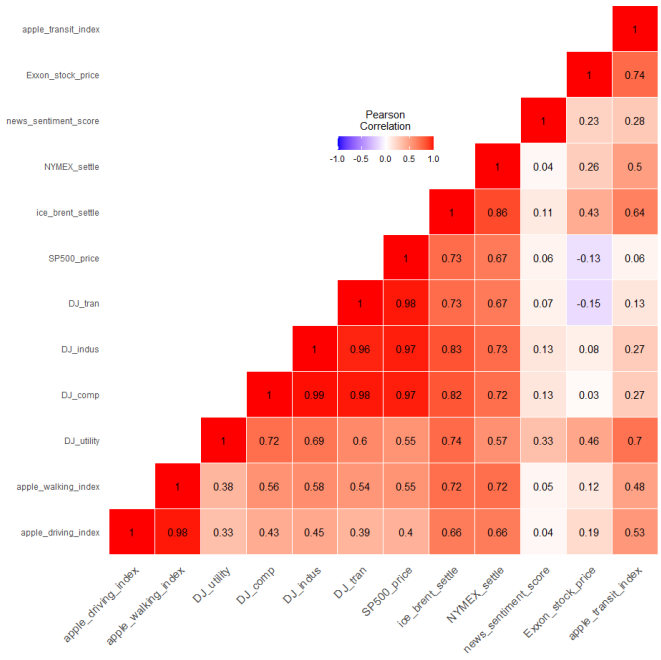


Figure 8

where x_i is the value of each observation, n is the total number of observations and z_i is the normalized value of each observation.

1.3.2. Regression Assumptions

For the purpose of inference, we decide to select a linear multiple regression model to estimate the stock price of Exxon. To proceed the model building, there are several assumptions of the data must be made.

The first assumption is that the relationship between the features and the target variable is linear. So one have:

$$Y = X \cdot \beta + \epsilon \tag{1.4}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{1.5}$$

The second assumption is that the error term ϵ is normally distributed. That is, $\epsilon \sim N(\mu, \sigma^2)$. One may verify this after fitting some linear models and plot the residuals' histogram.

The third assumption is that homoscedasticity of errors. That means, the σ^2 of ϵ should be constant. This assumption may also be easily checked after fitting some linear models and plot the residuals against date or other features.

The last assumption is that the residuals, the ϵ_i, need to be independent one to another. It is difficult to comprehensively verify this assumption. But, one may plot the residuals against time or date and see if

there are some obvious patterns.

1.3.3. Missing Values

The main missing data for the data set is the news sentiment. Many trading days have no news associated with Exxon and there are news coming out during the weekend or holiday. If one simply remove the dates with missing data, there are 81 observations left for the data set. Simply removing the missing data observation is not necessary a correct action to take. But, methods for imputation would also be problematic as well. Thus, for this part of the report, we decide to simply assume those missing news sentiments are missing at random (MAR) and simply remove those observations.

1.4. Model Selections

For this report, we considered two approaches to select the model. The first approach is the traditional both direction step-wise selection using AIC as the criterion. (This report does not include specific algorithm of step-wise selection since it is a very classical method; we assume the readers understand the mechanism of step-wise selection) We choose the direction to be "both", and begin with the full model including every single features. After the selection, below features are selected: Ice Brent Settle price, Dow Jones Composite, Dow Jones Industrial, Dow Jones Transportation, Dow Jones Utility and SP 500. The model including those 6 features has $R^2 = 0.9222$, which means the model has captured about 92% of variations of the stock prices. One can see more details from the output below:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.269e-14  3.223e-02   0.000 1.000000
ice_brent_settle  3.092e-01  7.462e-02   4.144 8.98e-05 ***
DJ_comp           3.379e+01  7.360e+00   4.591 1.77e-05 ***
DJ_indus         -1.441e+01  3.891e+00  -3.704 0.000406 ***
DJ_tran          -1.523e+01  2.933e+00  -5.194 1.75e-06 ***
DJ_utility       -3.820e+00  8.357e-01  -4.571 1.90e-05 ***
SP500_price      -2.093e+00  2.545e-01  -8.221 4.94e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.29 on 74 degrees of freedom
Multiple R-squared:  0.9222,Adjusted R-squared:  0.9159
F-statistic: 146.2 on 6 and 74 DF,  p-value: < 2.2e-16
```

Unfortunately, it does not include the feature news sentiment. To see if news sentiment can still capture some variation, we force the news sentiment stay in the model to run the AIC step-wise selection again. The 6 features selected previously still got selected, while the p-value of the news sentiment is 0.692049, which indicates it is not a significant predictor. The R^2 of the new model is 0.9224, which is also a very insignificant increase.

The second approach implemented is the step-wise selection using 10-fold cross validation's rooted mean squared prediction error as the criterion to select the variables. Using this criterion, only Ice Brent Settle price, Dow Jones Composite, Dow Jones Transportation, Dow Jones Utility and SP 500 are selected. This model still does not include the feature news sentiment. From the output below, one can see that it only has $R^2 = 0.7499$, while the SP 500 is no longer a significant predictor, with p-value $= 0.116$. This model lost a relatively large proportion of captured variation compared with the model selected by AIC; and the model makes SP 500 no longer significant. Therefore, for the sake of interpretability, we decide that the model selected by AIC is more appropriate to make inference about the relationship between stock price and other features.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -8.137e-16  5.701e-02   0.000    1.000
ice_brent_settle  8.316e-01  1.056e-01   7.879 1.88e-11 ***
DJ_tran          -1.484e+00  2.802e-01  -5.298 1.10e-06 ***
DJ_utility        4.941e-01  9.102e-02   5.428 6.55e-07 ***
SP500_price       4.406e-01  2.769e-01   1.591    0.116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5131 on 76 degrees of freedom
Multiple R-squared:  0.7499,Adjusted R-squared:  0.7367
F-statistic: 56.96 on 4 and 76 DF,  p-value: < 2.2e-16
```

1.5. Model Evaluation and Assumptions Check

From the Figure 9, one can see that the model selected by AIC indeed captured the trend of stock price change very well. Note that the y-axis of the figure is the normalized price of the Exxon's stock price.

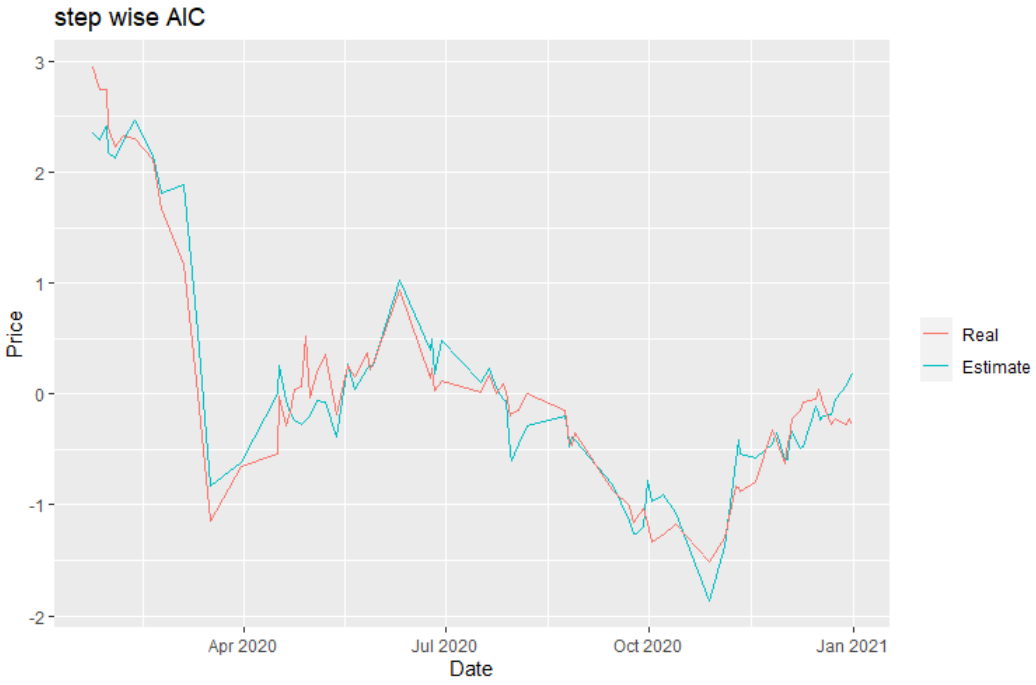


Figure 9: The red line is the real stock price of Exxon; the blue line is the estimated stock price using the model selected by step-wise method with AIC as criterion.

Even though the Figure 9 shows a nice capture of stock price variation, we still need to verify the model assumptions. One may use the shape of histogram to verify that whether the error term is following normal distributions. By the Figure 10, one can see that the residuals are roughly following a normal distribution. However, from the Figure 11, one can clearly see that the error is not evenly distributed around the line $y = 0$. This means the assumption of constant variance of term ϵ is not satisfied. Further, one can observe that the residuals are following a pattern of seasonal increase and decrease, especially for the first half of the year. This means the independence assumption of errors is also not hold. The reason for such pattern could be because there are some important features are missed. Another reason could be because the stock price may have some non-linear relationships with certain features. Moreover, it could be because the independence assumption of each observation is wrong.

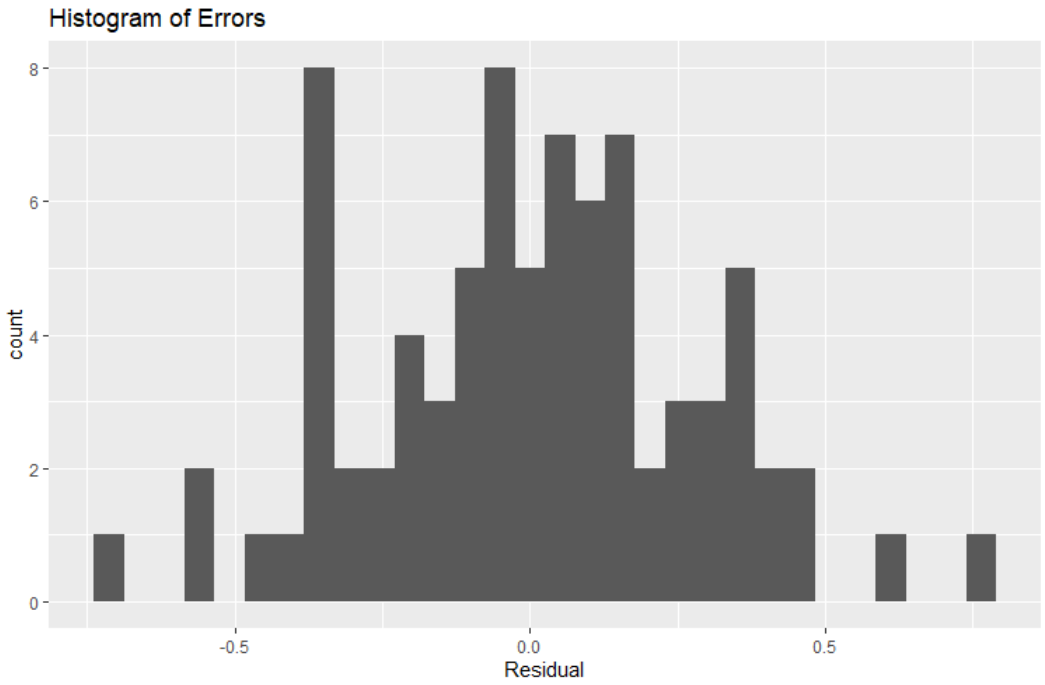


Figure 10

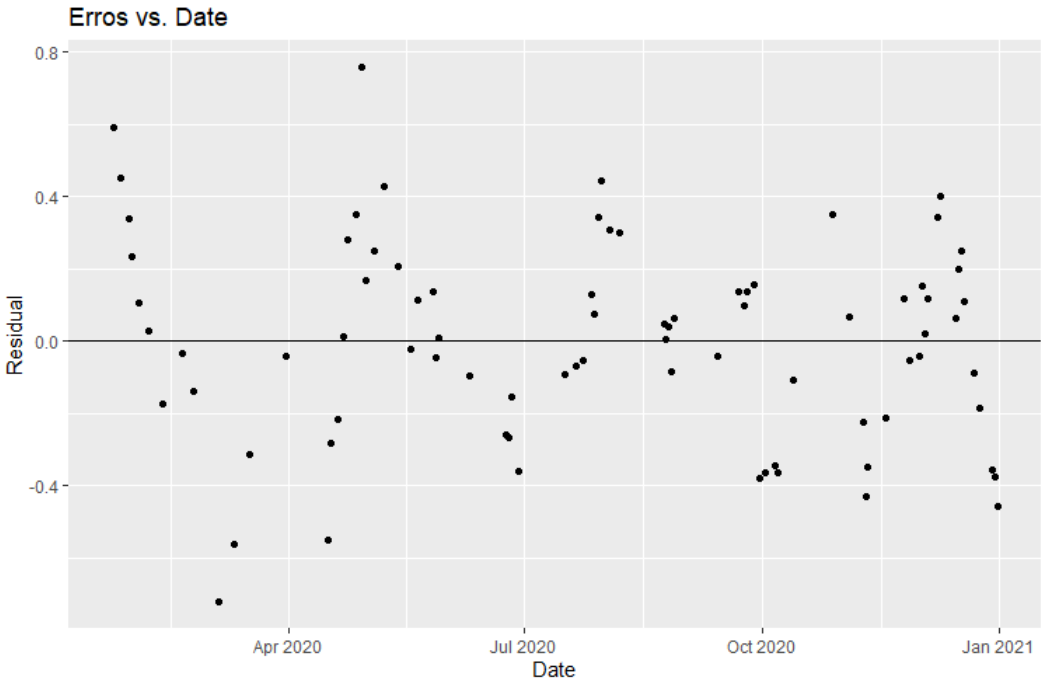


Figure 11

1.6. Summary of EDA and Linear Regression

From the above results, one can see that the Ice Brent, several Dow Jones Indexes and SP 500 are some very significant linear predictor for stock price of Exxon. However, by the investigation of the residuals of

the model, we find that the error is not having a constant variance; and the error term may not be independent to one another. We consider it is safe to conclude that the Crude Oil Commodity price and several major market indicators are linear correlated with the stock price of Exxon. However, there are some very interesting pattern of the variation did not captured by those features. Also, the news sentiment, the factor that we want to investigate, does not play a role here. Hence, we believe that the news sentiment may have a more complicated association with the stock price that is non-linear. To approach that, we decide to take advantage of Neural Network to track the association between the news sentiment and the stock price.

We performed similar methods to other energy companies and the results are very similar. Thus, the report does not include the results from those companies.

2. Neural Network Approach

In this section, we are devoted to exploiting how Recurrent Neural Network (RNN) [3] predicts stock price in a matured market, especially when additional features, such as news sentiment score, are counted into the model. The architecture of the neural network (NN) is chosen as Long short-term memory (LSTM) [2], which can not only process single data points but also entire sequences of data. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

Now we discuss the LSTM unit in details. Denote h_t as the hidden state vector and p_t as the input vector for time t. The forget gate f_t determines how much information of the cell sate c_{t-1} are discarded for time t, it accepts the output from the previous time step h_{t-1} and the new input p_t of the current time step. The input gate i_t determines how much the current time network input p_t is reserved into the new cell state c_t. It has three different components: 1) Get the state of the cell c_{t-1} that must be updated; 2) Create a new cell state $\hat{c}_t$; 3) Update the cell state to the current cell state c_t. The output gate o_t controls how much the newly created cell state c_t will be discarded. We have the following recursive formula for the forwad pass of an LSTM unit [2]:

$$\begin{aligned} f_t &= \sigma(W_f \cdot p_t + U_f \cdot h_{t-1} + b_f) \\ i_t &= \sigma(W_i \cdot p_t + U_i \cdot h_{t-1} + b_i) \\ o_t &= \sigma(W_o \cdot p_t + U_o \cdot h_{t-1} + b_o) \\ \hat{c}_t &= \tanh(W_c \cdot p_t + U_c \cdot h_{t-1} + b_c) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \hat{c}_t \\ h_t &= o_t \circ \tanh(c_t), \end{aligned}$$

where σ is a sigmoid function, $\circ$ is the element-wise multiplication and W_p, U_p and b_p are weight matrices and bias vector parameters which need to be learned during training, the subscript p can either be the input gate i, output gate o, the forget gate f or the memory cell c.

RNN is a type of artificial neural network with self-loop in its hidden layers, which enables RNN to use the previous state of the hidden neurons to learn the current state given the new input. The stock prices is a time series of length N, defined as $p_1, \cdots, p_N$, where p_i is the price of stock price on day $i, 1 \leq i \leq N$. Other features also can be considered as time series, for example, we can denote the news sentiment score on day i as a_i. In order to construct connection among the price and other features, we build a window X_i of a fixed size w at time i that each window contains all information from day i to day $i + w - 1$. For instance, the first window X_1 contains all information for the first w days, namely, $X_1 = (p_1, \cdots, p_w; a_1, \cdots, a_w)$. Our ultimate goal becomes to learn an approximate function $f(X_1, \cdots, X_t)$ and to use the hidden layer(s) learned by f to predict the last day's stock price p_{i+w-1} in window X_i. The common strategy to predict stock price using LSTM only consider one variable, namely the stock price itself, and one example see this post. To obtain a more accurate prediction, we propose a novel method that more features, for example news sentiment score, are concluded in LSTM.

2.1. Data preparation

Since the stock market only opens on weekdays, the prices freeze on weekends and we propagate the price on Friday forward to next Monday. We deal with other features similarly, such as DJ, S&P 500, as last valid

observation is forward to next valid backfill for missing data. We treat News Sentiment Score uniquely by filling with zero in missing values for daily input, which is natural that not every day has news. In this case, we believe the news sentiment has no contribution in stock price.

Moreover, we apply feature scaling [1], which normalizes the range of independent variables or features of data, to all input data for consistency. The values for sentiment scores are from -1 to 1 while the stock prices usually vary up to hundreds. To avoid the inaccuracy from scaling, we scale the range of all values in $[-1, 1]$ and forward scaled data in LSTM cell to learn.

As the stock price increases in time, bringing about the problem that most values in the test set are out of the scale of the train set and thus the model has to predict some numbers it has never seen before, which has a poor performance. In order to test the performance of our method based on RNN, we divide the set of windows into two part, $\mathcal{S}_1 = \{X_1, \cdots, X_d\}$ and $\mathcal{S}_2 = \{X_{d+1}, \cdots, X_{N-w+1}\}$, for training and testing, respectively. To investigate the learning ability with respect to training effectiveness, we explore the training ratio $\alpha := |\mathcal{S}_1|/(|\mathcal{S}_1| + |\mathcal{S}_2|)$ influence the mean square error (MSE) of the prediction, where $|\cdot|$ is the cardinality. Furthermore, we investigate how the size of the window w affect the prediction accuracy. It is obvious that the news have a bad, good or neutral influence on the related company's stocks in a short time and the impacts are reflected by the change of daily price. Does the news influence the stock price for the next day as well? We are interested the relation between the timeliness of the news sentiment and window size w, and try to figure out how the window size w influence the network learning.

2.2. Numerical Results of NN

We present the numerical prediction of Exxon Mobil's stock price based on RNN in this section. All data ranges from Jan. 2nd, 2020 to Dec. 31st, 2020. Some pre-treatment of data are achieved such as filling missing values and data scaling are described in Section 2.1. The stock price of Exxon Mobil and the related average daily news sentiment scores are available on WorldData.

Fig. 12 and Fig. 13 exhibit the time series of predicted prices and real prices of Exxon Mobil with various window sizes. From the predictions results, our method has a better prediction when a small window size w is selected. Overall, a network with larger window is more powerful and is able to learn more variables. However, a long window means that the system learn needs more inputs and is more difficult to train. For a larger size, it contains more parameters and some problems such as overfitting may occur.

Fig. 14 and Fig. 15 show the time series of prediction and the real prices of Exxon Mobil for various training ratios. The curves of these figures imply that our method can predict the stock price for a small value of α. For an extreme small training ratio, the model does not have enough data to learn and loses the ability of predicting stock prices.

Fig. 16 evaluates the performance of our method against training ratio for two different window sizes, $w = 2$ and $w = 7$. The performance score is defined as the square root of MSE, i.e. Score= $\sqrt{\sum_{i\in S} |p_i - \hat{p}_i|^2/|S|}$ where S is the underlying set and $\hat{p}_i$ is the predicted price for p_i. The test performance of $w = 2$ (left figure in Fig. 16) is better than the counterpart of $w = 7$ (right figure in Fig. 16), while the train score of $w = 7$ is smaller. Followed our previous discussion, there may be a reason behind that a larger window is more powerful but is weaker to predict in stock price because of overfitting. Since the competition is time-limited, we stop here and make some conjecture/guess why this happens. With the data using in this experiment, a smaller window size makes the prediction more accurate. The test score is very large and unacceptable when $\alpha < 0.2$, which implies that the learning process needs enough data. When the training ratio is large enough, our method can predict the stock price well and has a stable performance no matter how the training ratio increases.

2.3. Conclusion

We propose a novel stock-price-prediction method based on LSTM. Comparing the traditional strategy, we consider not only stock price data but also news sentiment scores. The window, as a basic input unite for GNN, contains all information in w days and the window size w affect the prediction of our method. A smaller size will lead to a better prediction in our case where the reason behind still needs to be explored. We conjecture that a larger window is more difficult to train well and may has overfitting. In general, our approach only needs around 20 percent of data where the length of data is one year.

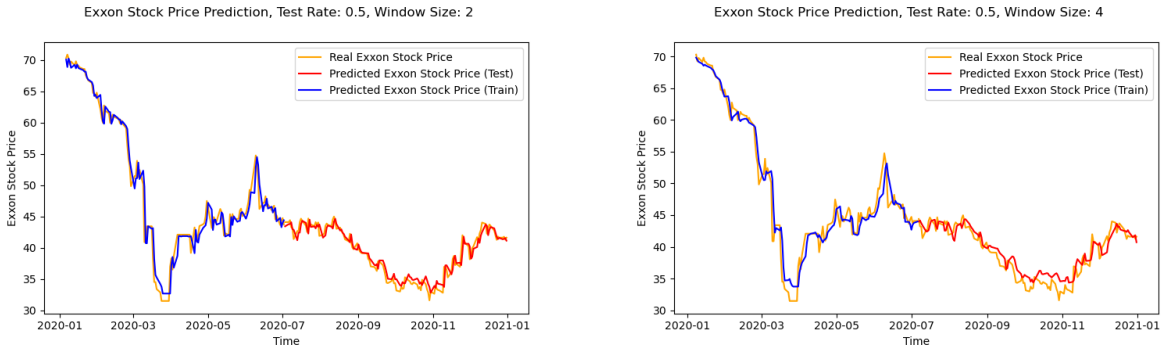


Figure 12: Time series of prediction and real stock prices with training ratio $\alpha = 0.5$. The window sizes are chosen as $w = 2$ (left) and as $w = 4$ (right). The orange curve is the real stock price of Exxon Mobil, the blue curve is the predicted price for training set and the red one is the predicted curve for testing set.

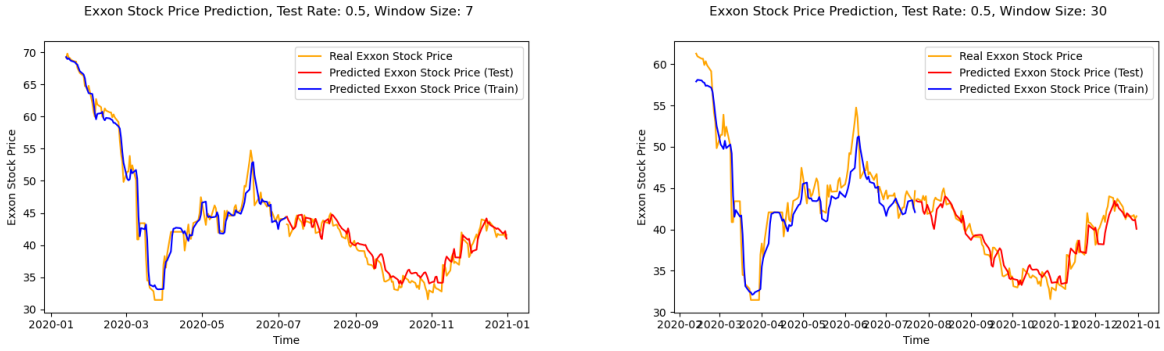


Figure 13: Time series of prediction and real stock prices with training ratio $\alpha = 0.5$. The window sizes are chosen as $w = 7$ (left) and as $w = 30$ (right). The orange curve is the real stock price of Exxon Mobil, the blue curve is the predicted price for training set and the red one is the predicted curve for testing set.

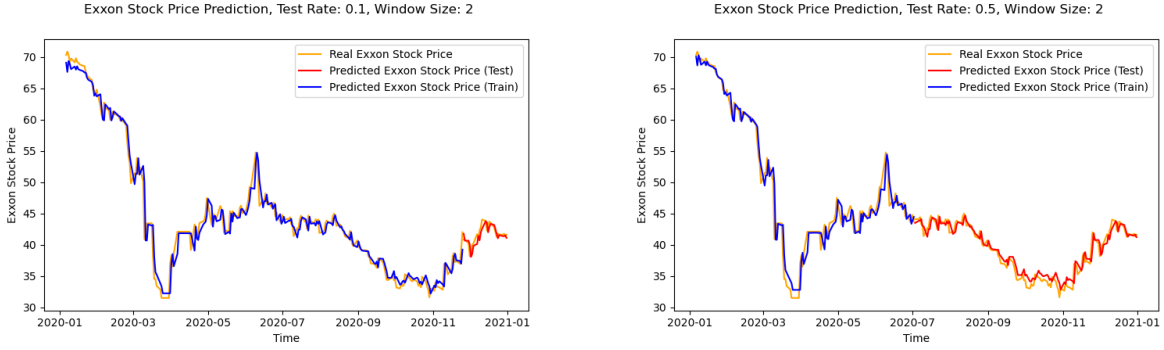


Figure 14: Time series of prediction and real stock prices with window size $w = 2$. The training ratios are chosen as $\alpha = 0.9$ (left) and as $\alpha = 0.5$ (right). The orange curve is the real stock price of Exxon Mobil, the blue curve is the predicted price for training set and the red one is the predicted curve for testing set.

3. Conclusion and Extension

In the first part of the report, we apply the linear regression to make some inference about the relationship between stock price and some of the features. We find features like crude oil commodity prices and major market indicator indexes are significant linear predictors. But we also find that the residuals of such models are still following some non-random patterns. So we decide to move on to Neural Network to investigate more.

A potential issue from the linear model approach is the news coming out during weekend or holiday.

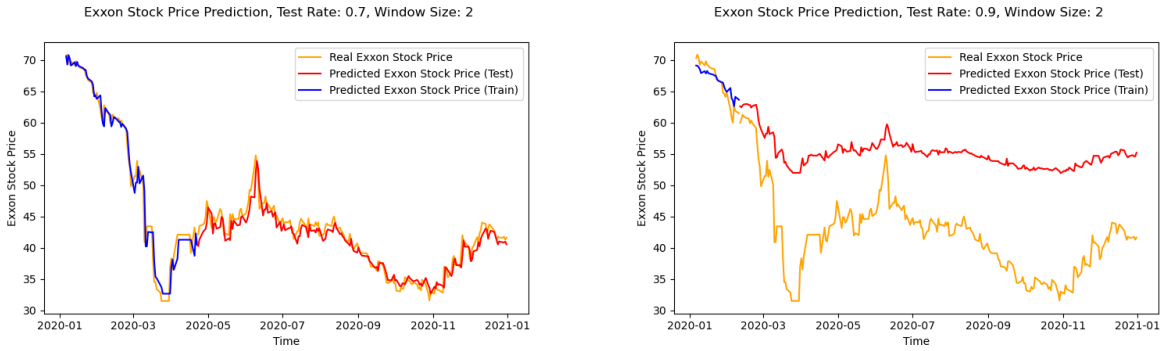


Figure 15: Time series of prediction and real stock prices with window size $w = 2$. The training ratios are chosen as $\alpha = 0.3$ (left) and as $\alpha = 0.1$ (right). The orange curve is the real stock price of Exxon Mobil, the blue curve is the predicted price for training set and the red one is the predicted curve for testing set.

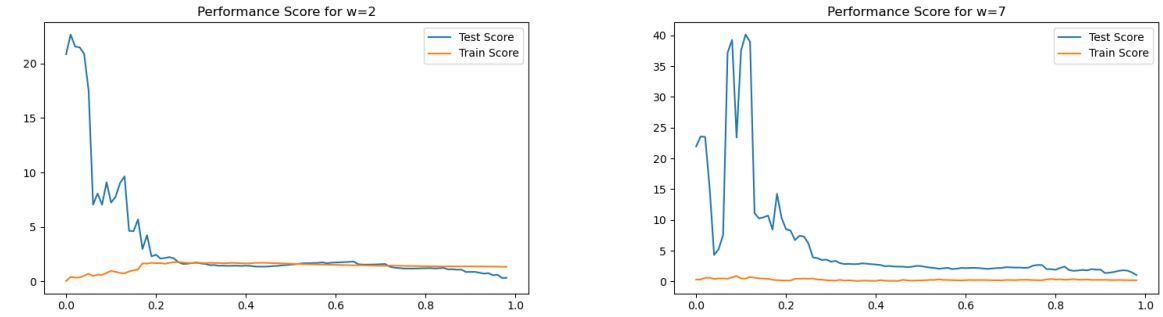


Figure 16: Performance score against training ratio with window size $w = 2$ (left) and $w = 7$ (right). The blue curve is the test score and the orange curve is the train score. The test score is unacceptable when $\alpha < 0.2$.

Those observations are discarded because their dates don't match with stock or commodity prices. However, those news certainly will create some impact on the market after the vacation. For future investigation, one may somehow add those news sentiment during vacation to the first trading day after the holiday. But, how to make such a modification is another subject deserved to be deeply studied.

References

[1] Joel Grus. Data science from scratch: first principles with python. *O'Reilly Media*, 2019.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[3] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.