Can Energy Market Infer Economic and Environmental Changes?

Abstract

In this paper, we explore the possibility of predicting gasoline price and carbon dioxide emission using crude oil price, as well as predicting gasoline price using crude oil futures price under the context of the United States. By examining classification and regression models, we implemented regular quantitative analysis methods and machine learning algorithms to find the best predictions. From classification models, high predictive power was observed in each scenario. However, in regression models, we only detected some predictive power for gasoline price but not for carbon dioxide emissions.

Keywords: Crude oil price, Carbon dioxide emissions, Energy futures, Optimization, Machine learning algorithms

1 Introduction

With the advancement of commercial and industrial operations, energy, especially crude oil, is undeniably the driving force for contemporary technical, economic and social practices. However, crude oil consumption contributed significantly to the emission of greenhouse gases, sparking growing universal concern about the environmental effects of carbon dioxide pollution from energy usage. Recognizing the correlation between greenhouse gas emissions and the drivers of both economic and social activities thus the nature of air quality attracts huge amounts of interest from researchers and the government. This article focuses on researching the causal relationship among crude oil price changes, energy consumption fluctuations, and carbon dioxide emissions, as well as testing whether crude oil price could accurately predict greenhouse gas emission changes in the following couple of weeks. To illustrate the predictive power of crude oil price and other economic and environmental variables, in section 2 of the paper, we demonstrate the influential factors by qualitative analysis and our data processing steps. In section 3, we apply quantitative analysis using different models, in order to see the predictive power and relative error. In section 4, we discuss some tangent issues and possible improvements. Finally, in section 5, we present our conclusion and key takeaways from this academic study.

2 Data Analysis and Processing

2.1 Qualitative Analysis

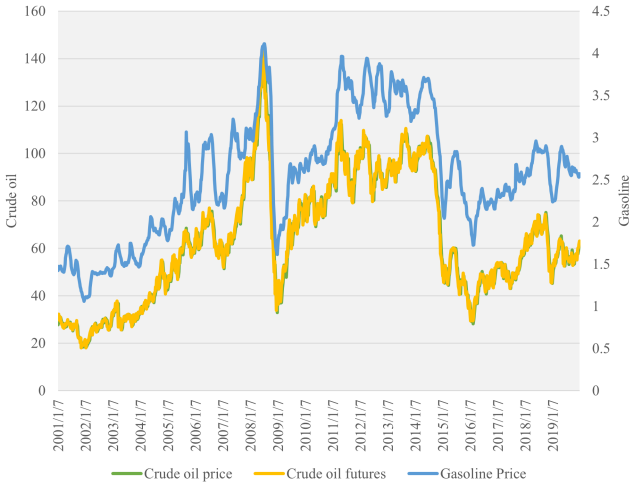

Figure 1. Crude oil price and futures, Gasoline price (2001 Jan - 2019 Jan).

In this paper, we try to examine how crude price and crude futures price affect gasoline price and how crude price affects the concentration of carbon dioxide in the atmosphere. From a historical standpoint, crude price and gasoline price followed a similar pattern in light of the fact that gasoline is produced from crude oil (U.S. Energy Information Administration, n.d.). In addition, the production of crude oil, along with the consumption of gasoline, also contributes to the increasing concentration of carbon dioxide in the environment (United States Environmental Protection Agency, n.d.). In figure 1, we showed a plot of crude oil price and futures and gasoline price that suggests they have similar trends. In the context of the United States, we derived three fundamental questions for our research:

1. Can crude oil futures price predict gasoline price?
2. Can crude oil price predict gasoline price?
3. Can crude oil price predict carbon dioxide emission?

However, when it comes to gasoline price and carbon dioxide, the crude price and crude futures price are not the only effective factors. Taking the emission of carbon dioxide as an example: the consumption of coal and natural gas can increase the level of greenhouse gas release. The growing population in the United States also indicates the expanding need for transportation, electricity, etc, resulting in more carbon emissions as well. Therefore, we included additional explanatory variables in our models for the purpose of predicting the response variables more accurately. Additional variables that we employed for different time periods are listed as a reference below in section 2.2 Quantitative Analysis.

2.2 Quantitative Analysis

First, we present a list of variables we have used in our analysis in Table 2. The Y in each cell represents the usage of this variable in this question, this time period. We explained our reasoning in the following sections.

2.2.1 Basic data

To examine the three fundamental questions, we collected weekly data on historical crude price, Crude Oil WTI Futures price (expiring Mar 21), gasoline price, as well as the weekly average concentration of carbon dioxide at Mauna Loa.

2.2.2 Additional data

In addition to the basic data, other explanatory variables that may influence gasoline price and carbon dioxide concentration were added. For gasoline price changes, we additionally utilized the Consumer Price Index (CPI), the U.S. population, amount of public roads in the U.S., urban population, crude oil production, the total sale of automobiles in the U.S. seasonally adjusted, average state fuel taxes, and U.S. real GDP. To elucidate carbon dioxide changes, we further employed Consumer Price Index (CPI), coal consumption, natural gasoline consumption, the U.S. population, Tree cover loss in the U.S., and U.S. real GDP. Table 2 exhibits an overview of the data. For gasoline price changes, crude futures price versus gasoline price and crude price versus gasoline price employ identical factors since we assume they have the same influence factors.

2.3 Data Processing

In this section, we discussed the data processing part. The time spans and frequencies of the main data were proper for this research. However, the additional data, whose time spans and frequencies were different, made data preprocessing an essential step. In this section, we talked about how we resampled our data by adjusting data periods and data frequency, created explanatory and response variables, and preprocessed data by normalization and PCA. The data processing procedure was done in python with the aid of excel.

2.3.1 Data cleaning

Firstly, we examined the time period of all the data available and found it necessary to divide our investigation into 3 possible periods, in order to coordinate the duration of datasets and make the sample size as large as possible. The final periods are as follows:

1. **Whole period**: 2001-2019, including in total 18 years of weekly data.
2. **Period 1**: 2001-2009, including data pre-global financial crisis.
3. **Period 2**: 2010-2019, including data post-global financial crisis, also more additional variables

Data cleaning was conducted in the exact same manner during these three time periods because 2009 is the year of the financial crisis, thus we presumed possible different behaviors for the scenarios that attracted attention, and a portion of the additional variables' data is only available after 2010. Since the start times were late, we dropped the two datasets of electrical vehicle ETF and average state fuel taxes and the discussion of them has been moved to section 4 Discussion.

Secondly, we upsampled our additional variables. Out of 14 additional datasets, there were five datasets

Questions		Crude oil price predict gasoline price & Crude oil futures predict gasoline price			Crude oil price predict CO2		
Variables		Whole period	Period 1	Period 2	Whole period	Period 1	Period 2
Basic variables	CO2				Y	Y	Y
	Gasoline price	Y	Y	Y			
	Crude oil price	Y	Y	Y	Y	Y	Y
	Crude oil futures price						
Additional variables	CPI	Y	Y	Y	Y	Y	Y
	Coal Consumption			Y	Y	Y	
	Natural Gas Consumption			Y	Y	Y	
	US population	Y	Y	Y	Y	Y	Y
	Public Roads	Y	Y	Y			
	US urban population	Y	Y	Y			
	Crude oil production	Y	Y	Y	Y	Y	Y
	Dow Jones U.S. Automobiles & Parts Index			Y			
	Total sale of automobiles in USA seasonally adjusted			Y			
	S&P GSCI Live Cattle Index Spot						Y
	Tree cover loss in the US				Y	Y	Y
	USD Mexican exchange rate			Y			Y
	US real GDP	Y	Y	Y	Y	Y	Y

Figure 2. Basic variables and additional variables.

with monthly frequency, three datasets with yearly frequency, and one dataset with quarterly frequency. For prices, consumption, tree cover loss, and GDP, an upsampling variation method was applied. For population and roads, which generally had incremental trends, the step method was applied. We adjusted the variation to weekly by multiplying the root of the number of weeks in the corresponding period, then normally distributed noises by creating zero mean and weekly standard deviation. These noises were added to the original period data. In this way, weekly data with relative variation were successfully produced.

After matching up the duration and frequency we proceeded to create variables for further usage to our models.

2.3.2 Creating explanatory and response variables

Firstly, the explanatory variables for basic scenarios, including crude oil futures and crude oil price, were generated. To know how long it would take for explanatory variables to have impacts on a particular response variable, we calculated percent changes of 1, 2, 3, 5, and 10 weeks from the original variables to conduct feature engineering. On the other hand, while adding additional explanatory variables in the models, only the original values of those variables were employed.

Secondly, we generated the response variables, including gasoline price and carbon dioxide emissions. For the regression models, we took the one-week percent change of the dependent variables of each question as Y. For classification models, we utilized one-hot encoding and marked one-week incremental change to be 1 and decremental change to be 0 as Y. It is worth saying that we generated backward historical change features for the basic variables.

2.3.3 Feature engineering

After splitting the training, validation, and testing sets, data were preprocessed by normalization and PCA. In data normalization, the mean of the training dataset was extracted from both the training set and testing set, and they were both divided by the standard deviation of the training set. With regard to PCA, the number of variables of our models was not large enough, thus we set it to explain 95% of the variation.

Then we proceeded to apply our models in python.

3 Modeling

3.1 Models Introduction

We employed regression models and classification models using machine learning algorithms. Our implementation of modeling was based on python. The results derived from the classification models exhibited the predictive power of our independent variables in the direction of change independent variables. On the other hand, the results from the regression models represented the predictive power on the dimension, the amount changed, of the dependent variable. We assumed that while some variables may be able to predict or influence the dimension, they could not predict whether the price change was positive or negative. Below are the models we have implemented and their main concepts.

3.2 Results

In the core of our investigation of the problems, we compared our results of predictive power in the aspect of time and function. Time comparison was in the following manner:

1. **Comparison 1**:Compare the predictive power in the whole period (2001-2019) of basic variables with and without additional variables. This aims to conclude that adding additional variables can improve the basic variables' predictive power with the same duration and start time.

2. **Comparison 2**: Compare the predictive power of the same variables in the same duration (period 1 2001-2009 and period 2 2010-2019) This aims to conclude that whether different time periods can influence the ability to predict with the same sets of basic and additional variables.

3. **Comparison 3**: Compare the predictive power in the same period (period 2 2010-2019) with extra additional variables than period 1. This intends to conclude whether these extra new variables can improve our prediction with the same duration and start time. Since these new variables only include data from a late start time, we can expect their future assistance in this field.

The first comparison aims to demonstrate whether the additional independent variable can improve the predictive power in terms of the basic independent variables models in the long run. The last two comparisons contribute to the short-term period examination as well as the elimination of the global financial crisis effects.

There were also two functions of predictions, forecasting if the change was upwards or downwards (direction) and forecasting how large the change was (dimension). The function of prediction was evaluated by our classification models and regression models. The functional comparison was in the following manner:

1. **Classification** models evaluate the directional prediction and give the accuracy score by the probability of correct prediction of change.

2. **Regression** models evaluate the dimensional prediction and give the proportion of the data explained, which is the R-squared value.

We view the scores returned from every model with equal importance thus we rely on the mean of the scores to determine the strength of predictive powers. Classification results were compared to 0.5 as in no effect and regression results were compared to 0 as in no effect.

Then we compared the results of our 3 main questions according to the above-mentioned metrics. We presented the detailed result for the pair of models in comparison 1 and the comparison result only for the other two comparisons. The rest of our detailed results can be found in the appendix.

3.3 Can crude oil futures price predict gasoline price?

To dig into the relationship between crude oil futures price changes and gasoline price fluctuations, we trained both regression models and classification models.

By taking the average score of these models for the whole period including only basic variables, we noticed that the classifier models had an average of

Models	Concept
Multiple Linear Regression	It aims to model the relationship between the explanatory variables and a response variable by fitting a linear equation to the observed data.
Random Forest	It fits a number of classifying decision trees as an ensemble for on sub-samples of the dataset and uses averaging to control overfitting (“Random Forests”, n.d.).
Gradient Boosting	It creates a prediction model in the form of an ensemble of weak models, typically decision trees, to minimize the error of the overall model (“Time series forecast with stochastic gradient boosting”, n.d.).
XGBoost	It optimizes every loss function, including logistic and pairwise ranking, to improve the random forests model (“XGBoost Documentation”, n.d.).
AdaBoost	It fits a regressor on the original dataset and then fits additional copies but the weights of instances are adjusted according to the error of the current prediction. Therefore, subsequent regressors focus more on difficult cases (“Adaboost”, n.d.).
Logistic Regression	It is used to determine if an independent variable has an effect on a binary dependent variable (“Logistic Regression”, n.d.).
Gaussian Naïve Bayes	It applying Bayes’ theorem with the assumption of conditional independence between every pair of features. The likelihood (distribution) of the features is assumed to be Gaussian (“Gaussian Naive Bayes”, n.d.).
Support Vector Classification	It is based on libsvm, constructs a set of hyper-planes in a high or infinite dimensional space, optimized by finding that has the largest distance to the nearest training data points of any class (functional margin). Three kernels, including linear, polynomial, and Radial Basis Function are implemented in this project (“Support Vector Machines”, n.d.).
Neuro Network	Given a set of features and a target, it can learn a non-linear function approximator for classification. Each neuron in the hidden layer transform the weighted value form the previous layer (“Neural Network Models (supervised)”, n.d.).

Table 1. Model Description

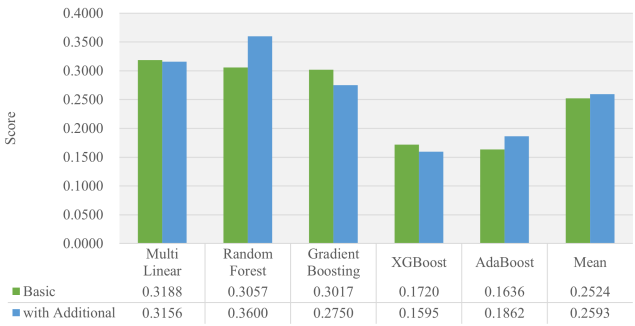

Figure 3. Crude oil futures price predict gasoline price, regression results.

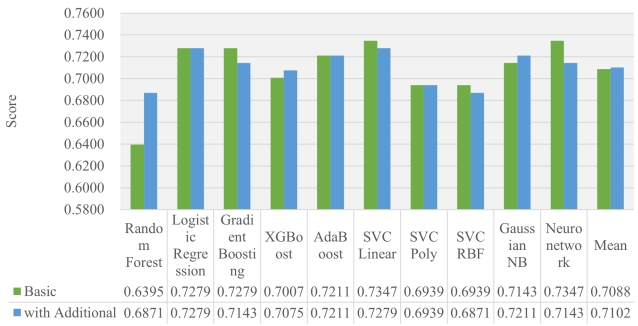

Figure 4. Crude oil futures price predict gasoline price, classification results.

0.7088, showing that the predictive power of crude oil futures was relatively high and significant for gasoline price changes. However, the regressor models had an average of 0.2524, mirroring that regressor did not have many contributions to offer convincing predictions of the relationship.

Considering additional explanatory variables, the results were slightly improved. As for the regression models, the average accuracy score of those models was around 0.2593, which performed a bit better than regression models with only basic variables but had not much predictive power as well. Regarding the classification models, the average accuracy score of them was about 0.7102, which was higher than that computed from the fundamental scenario. Thus, as shown in Table 2 and Table 3, comparison 1, most of the models would generate higher scores in the long run when additional variables were included.

After splitting the entire time span into two periods, we noticed that, in Table 3, comparison 2, the

Regression comparison results			
Regression models	**Comparison 1**	**Comparison 2**	**Comparison 3**
Multi Linear	-0.0032	0.1440	0.0003
Random Forest	0.0543	0.1319	0.0729
Gradient Boosting	-0.0267	0.0961	-0.0201
XGBoost	-0.0125	0.1011	0.0999
AdaBoost	0.0226	0.0118	0.1755
Mean	0.0069	0.0970	0.0657

Table 2. Crude oil futures price predict gasoline price, regression comparison results.

Classification comparison results			
Classification models	**Comparison 1**	**Comparison 2**	**Comparison 3**
Random Forest	0.0476	0.0405	0.0824
Logistic Regression	0.0000	0.1090	0.0118
Gradient Boosting	-0.0136	0.1813	-0.0118
XGBoost	0.0068	0.1157	-0.0118
AdaBoost	0.0000	0.0523	-0.0118
SVC Linear	-0.0068	0.0854	0.0235
SVC Poly	0.0000	0.0993	0.0118
SVC RBF	-0.0068	0.2027	-0.0118
Gaussian NB	0.0068	0.0901	-0.0118
Neuro network	-0.0204	0.2872	-0.1059
Mean	0.0014	0.1263	-0.0035

Table 3. Crude oil futures price predict gasoline price, classification comparison results.

average score of the second period derived from the classifiers was higher than those of the first period when these two periods have the same additional variables. However, as for Table 3, comparison 3, the mean results derived from classifiers within period 2 performed better than those within period 1 because period 2 includes two more variables than period 1, which are Dow Jones U.S. Automobiles & Parts Index and USD Mexican exchange rate. Thus, we can show that these two variables could successfully improve only the dimensional predictive power for tracking the gasoline price fluctuations.

Then, by closely investigating both the regression scores and classification scores of these two periods, we observed that period 1 had lower accuracy scores than period 2 when crude oil futures prices were increasing within period 1. As for period 2, crude oil futures prices were decreasing most of the time with relatively higher accuracy scores. Therefore, it exhibits that there exist different responses for rising crude prices and falling crude prices, which could be a result of the financial crisis. The logic behind this could be explained by our conjectures such that when crude oil's futures price rise, the market has a higher demand for crude oil, which attracts more additional variables to contribute to overall predictive power, so that those explanatory variables we selected may have relatively less predictive power.

3.4 Can crude oil price predict gasoline price?

We have done a similar process to investigate the predictive power of crude price on gasoline price.

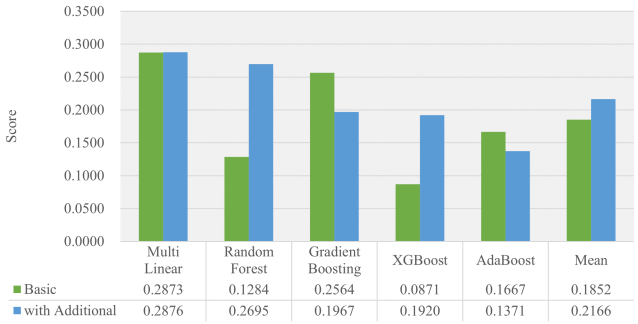

Figure 5. Crude oil price predict gasoline price, regression results.

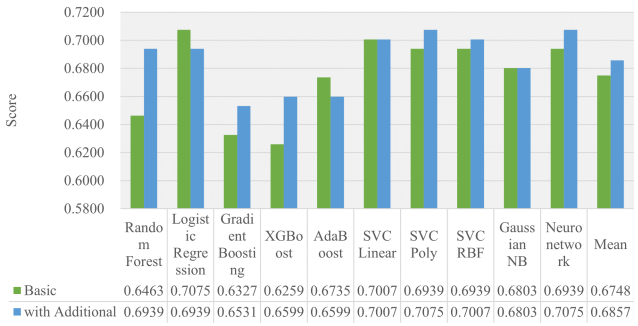

Figure 6. Crude oil price predict gasoline price, classification results.

Firstly, we analyzed the relationship between the basic variables: crude price and gasoline price. The classification results show that crude price has strong predictive power on gasoline price change's direction because all models returned an average accuracy score of 0.6748, which is well above 0.5. The regression results exhibit that crude price has an average of

Regression comparison results			
Regression models	**Comparison 1**	**Comparison 2**	**Comparison 3**
Multi Linear	0.0003	0.1015	0.0116
Random Forest	0.1411	0.3334	0.0176
Gradient Boosting	-0.0598	0.5279	0.0050
XGBoost	0.1049	0.4332	0.0306
AdaBoost	-0.0295	-0.0433	-0.0219
Mean	0.0314	0.2705	0.0086

Table 4. Crude oil price predict gasoline price, regression comparison results.

Classification comparison results			
Classification models	**Comparison 1**	**Comparison 2**	**Comparison 3**
Random Forest	0.0476	0.0316	0.0235
Logistic Regression	-0.0136	0.0598	0.0118
Gradient Boosting	0.0204	0.0598	-0.0235
XGBoost	0.0340	0.1207	0.0118
AdaBoost	-0.0136	0.1325	-0.0471
SVC Linear	0.0000	0.0737	0.0118
SVC Poly	0.0136	0.0199	0.0000
SVC RBF	0.0068	0.0854	0.0118
Gaussian NB	0.0000	0.0762	-0.0118
Neuro network	0.0136	0.1418	0.0118
Mean	0.0109	0.0801	0.0000

Table 5. Crude oil price predict gasoline price, classification comparison results.

0.1852, which means the crude price has low predictive power in the dimension of gasoline price change. However, in comparison to crude futures price, crude price's accuracy score is generally lower by about 0.05, meaning crude oil price has lower power than its futures.

Then we added additional variables. Considering the whole time period from 2001 – 2019, in the aspect of predicting the direction of gasoline price change, there was no significant improvement with the average model accuracy score according to our results in comparison 1. Both classification models and regression models had changes below 0.05, as shown in Table 5 and Table 4, comparison 1.

Then we divided the investigation period into two segments, 2001-2009 and 2010-2019. During period 1, we kept the additional variables the same as when we conducted the whole time period and there was a significant enhancement in several models but not all for both classification models and regression models. Regression models overall were improved by 0.2705 and classification models only 0.0801, as shown in Table 5 and Table 4, comparison 2. The before mentioned upward and downward trend also exhibits in crude oil price. We can conclude that after the financial crisis, the crude oil price is more closely related to the amount of change in gasoline price. Similar to the comparison 3 results in the last question, Table 4 Comparison 3 shows the same extra additional variables did not make significant changes to either of the predictive power.

In summary, the crude price itself does not have much predictive power in gasoline price change's dimension but is significantly important for predicting gasoline price change's direction. Also, both predictive powers have been justified to be weaker than crude oil futures price. The results also indicate that adding the additional variables for the whole period, which were the same as in period 1, cannot improve any of the predictive powers. On the contrary, after the financial crisis, the predictive powers both upgraded.

3.5 Can crude oil price predict carbon dioxide emission?

Carbon dioxide emission was examined in the same way but with different sets of additional variables.

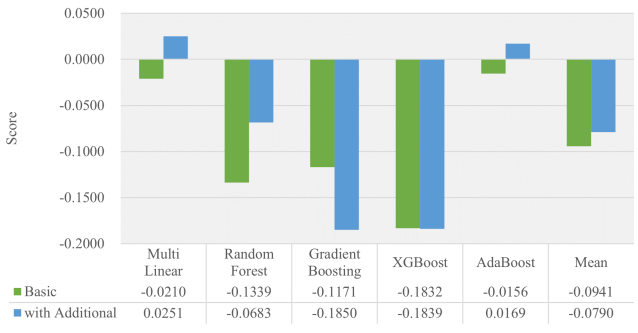

Figure 7. Crude oil price predict carbon dioxide emission, regression results.

With only basic data, we found that the scores were relatively low for both regression and classification models, meaning that oil futures price alone has no predictive power on neither the quantity nor the trend of the carbon emission. The average classification score was 0.5088 while the average regres-

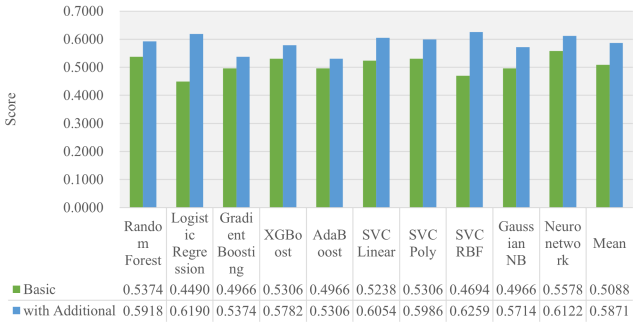

Figure 8. Crude oil price predict carbon dioxide emission, classification results.

Regression comparison results			
Regression models	**Comparison 1**	**Comparison 2**	**Comparison 3**
Multi Linear	0.0461	0.0256	-0.0106
Random Forest	0.0655	-0.1609	0.0450
Gradient Boosting	-0.0679	-0.2399	0.0300
XGBoost	-0.0007	-0.3177	0.0106
AdaBoost	0.0325	-0.0156	-0.0098
Mean	0.0151	-0.1417	0.0131

Table 6. Crude oil price predict carbon dioxide emission, regression comparison results.

Classification comparison results			
Classification models	**Comparison 1**	**Comparison 2**	**Comparison 3**
Random Forest	0.0544	-0.0532	0.0706
Logistic Regression	0.1701	0.1931	-0.0235
Gradient Boosting	0.0408	-0.0793	-0.0235
XGBoost	0.0476	-0.0957	-0.0824
AdaBoost	0.0340	0.0052	0.0235
SVC Linear	0.0816	0.1203	0.0353
SVC Poly	0.0680	0.1859	-0.0235
SVC RBF	0.1565	0.1275	-0.0588
Gaussian NB	0.0748	-0.0062	0.0471
Neuro network	0.0544	0.0519	0.1059
Mean	0.0782	0.0450	0.0071

Table 7. Crude oil price predict carbon dioxide emission, classification comparison results.

sion score was -0.0941. In addition to the crude futures price, we then added other explanatory variables to predict the carbon emission. From 2001 to 2019, with those additional variables, we observed that the average accuracy scores improved by 0.0782 for classification models as we can see from classification comparison results, as shown in Table 7, comparison 1. However, there was no significant improvement in regression scores. Moreover, we split our models into two-time spans to comprise more explanatory variables. From 2001-2009, we noticed a slight increase in the scores for most of the regression and classification models compared with the models with only basic variables. From 2010-2019, with two more explanatory variables (S&P GSCI Live Cattle Index and the exchange rate between USD and Mexican Peso), the classification models yielded better results, though the accuracy scores were still below 0.8, as shown in Table 8. Still, a significant improvement in the regression scores was not investigated. To examine the impact of the two additional explanatory variables from 2009 January to 2019 December, we tried to eliminate the variables. In regression comparison results - comparison 2, we found out that 80% of the classification scores were lower than those of the first period: 2001 January to 2008 December. In other words, for most of the regression models, the forecasting power of the independent variables is higher from 2001 to 2009 than from 2010 to 2019. In general, when we compared the models with additional variables to the models with only oil futures as an independent variable, we built up the conclusion that the crude futures price along with other explanatory variables can predict the trend of CO_2 with an average of 0.5871 accuracy score. However, those variables could not be used to predict the quantitative change of CO_2. We believe that the carbon emission in the atmosphere is affected by countless factors and is changing constantly. Thus, it is difficult to use a particular model to predict the quantitative change of CO_2 concentration, but it is possible to forecast the trend of carbon emission with our models.

4 Discussion

4.1 Additional time lag investigation

As for how we decided the response variables time lag should equal to one week, we did extra research with alteration of the period parameter in the percentage change function. However, this topic is tangent to our main conclusion thus was not mentioned in our former sections. Intuitively, since the percentage change does not exist for the first row if we set

the period parameter to one, we can discard the first row in the data set or move up the percentage change data to fill the void. After careful comparison, we examined that, setting the response variables' percent change to within period 1 and discarding the void rows produced the best results. In general, it takes about one week for the explanatory variables to affect the response variables.

4.2 Possible improvements

There are still a few features that can be improved through further exploration and research.

- **Unusable models.** By applying only basic variables, XgBoost and AdaBoost generally cannot predict extreme prices at an accurate level since they typically returned negative scores.
- **Extreme value interference.** At the most extreme values, there were gaps between actual values and predicted values. This feature may be improved by certain deep learning methods which can predict extreme events in turbulent dynamical systems, or in a simpler way, eliminating them from the datasets(Majda, 2020). These two methods are both worthy to explore further to improve fitting results.
- **Time lag.** In addition, features with time considered were only created for the basic variables of each fundamental question and not its additional variables. It might need more effort to set different time lags for every variable, since different variables may take different time durations to influence the dependent variables. This progress might be complex and fussy to handle, but still worth exploring to improve the results.

4.3 Conjectures

- **Financial crisis.** As we mentioned before in section 3.5, we found for 80% of the models that have been implemented, the forecasting power of the independent variables was lower in period 2 (post-financial crisis). Before 2009 January, the crude futures prices presented a gradually rising trend. However, it started fluctuating without a pattern after 2009 January. Our speculation is that the oil futures prices lost their trend after the financial crisis in 2008. Thus, the predictive power became lower in the second period.
- **Classification results were generally higher.** In our findings, most of the results were higher in classification models and low in regression models. We suggest this phenomenon is due to classification models having a simpler dependent variable. In other words, it is harder to predict a specific number than just predicting a boolean value.
- **Taxes and Electrical vehicle development.** We initially included these two variables as additional variables but their time durations were not long enough to fit in our models. However, we surmise that taxes on gasoline and electric vehicle development will decrease the demand for gasoline, consequently decreasing the price for gasoline.

5 Conclusion

In conclusion, by conducting machine learning methods, we have found significant predictive power in the direction of change between crude oil futures price and gasoline price, crude oil price, and gasoline price, but not as significant for carbon dioxide emission. Likewise, for predictive power on the dimension of change, we have only observed significant predictive power for the first two questions but not for greenhouse gas emission.

Also, with the assistance of additional explanatory variables, we were able to improve the predictive power of our basic variables. In the aspect of time, we have concluded that with the same duration, same start time, adding additional variables improved the predictive power of basic variables. With the same duration, same variables, different start times only had significant improvement in using crude oil price to predict gasoline price. Thus, the financial crisis only impacted the relationship between crude oil price and gasoline price. With the same duration and same start time, the extra additional variables slightly improved the prediction results. On the other hand, the functional comparison indicated

that classification scores were generally higher than regression scores, suggesting that the prediction of the direction of change is more accessible than a prediction of the dimension of change. Now that there is a larger amount of available data within the post-crisis period due to the global economic recovery, we can expect better forecasting powers for crude oil for gasoline prices and emission for carbon dioxide.

References

Adaboost. (n.d.). https://scikit-learn.org/stable/modules/ensemble.html#adaboost (accessed: 02.20.2021)

Gaussian naive bayes. (n.d.). https://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes (accessed: 02.20.2021)

Logistic regression. (n.d.). https://www.statisticssolutions.com/regression-analysis-logistic-regression/ (accessed: 02.20.2021)

Majda, D. Q. A. J. (2020). Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, *177*(1), 52–59. https://doi.org/10.1073/pnas.1917285117

Neural network models (supervised). (n.d.). https://scikit-learn.org/stable/modules/neural_networks_supervised.html (accessed: 02.20.2021)

Random forests. (n.d.). https://scikit-learn.org/stable/modules/neural_networks_supervised.html (accessed: 02.20.2021)

Support vector machines. (n.d.). https://scikit-learn.org/stable/modules/svm.html# (accessed: 02.20.2021)

Time series forecast with stochastic gradient boosting. (n.d.). https://rstudio-pubs-static.s3.amazonaws.com/161075_05ce98dc51c844e0833c06835c9ce4c3.html (accessed: 02.20.2021)

United States Environmental Protection Agency. (n.d.). *Carbon dioxide emissions*. https://www.epa.gov/ghgemissions/overview-greenhouse-gases#carbon-dioxide (accessed: 02.25.2021)

U.S. Energy Information Administration. (n.d.). *Oil and petroleum products explained*. https://www.eia.gov/energyexplained/oil-and-petroleum-products/ (accessed: 02.25.2021)

Xgboost documentation. (n.d.). https://xgboost.readthedocs.io/en/latest/index.html (accessed: 02.20.2021)

Appendix

Classification models	Basic	with Additional	Period 1	Period 2 no extra	Period 2 with extra
Random Forest	0.6395	0.6871	0.6066	0.6471	0.7294
Logistic Regression	0.7279	0.7279	0.6557	0.7647	0.7765
Gradient Boosting	0.7279	0.7143	0.5246	0.7059	0.6941
XGBoost	0.7007	0.7075	0.5902	0.7059	0.6941
AdaBoost	0.7211	0.7211	0.6066	0.6588	0.6471
SVC Linear	0.7347	0.7279	0.6557	0.7412	0.7647
SVC Poly	0.6939	0.6939	0.6066	0.7059	0.7176
SVC RBF	0.6939	0.6871	0.5738	0.7765	0.7647
Gaussian NB	0.7143	0.7211	0.6393	0.7294	0.7176
Neuro network	0.7347	0.7143	0.5246	0.8118	0.7059
Mean	0.7088	0.7102	0.5984	0.7247	0.7212

Figure 9. Crude oil futures price predict gasoline price, detailed classification results.

Regression models	Basic	with Additional	Period 1	Period 2 no extra	Period 2 with extra
Multi Linear	0.3188	0.3156	0.2881	0.4322	0.4324
Random Forest	0.3057	0.3600	0.1780	0.3098	0.3828
Gradient Boosting	0.3017	0.2750	0.2586	0.3547	0.3346
XGBoost	0.1720	0.1595	0.1232	0.2243	0.3242
AdaBoost	0.1636	0.1862	0.1326	0.1444	0.3199
Mean	0.2524	0.2593	0.1961	0.2931	0.3588

Figure 10. Crude oil futures price predict gasoline price, detailed regression results.

Classification models	Basic	with Additional	Period 1	Period 2 no extra	Period 2 with extra
Random Forest	0.6463	0.6939	0.7213	0.7529	0.7765
Logistic Regression	0.7075	0.6939	0.7049	0.7647	0.7765
Gradient Boosting	0.6327	0.6531	0.7049	0.7647	0.7412
XGBoost	0.6259	0.6599	0.6557	0.7765	0.7882
AdaBoost	0.6735	0.6599	0.6557	0.7882	0.7412
SVC Linear	0.7007	0.7007	0.6557	0.7294	0.7412
SVC Poly	0.6939	0.7075	0.7213	0.7412	0.7412
SVC RBF	0.6939	0.7007	0.6557	0.7412	0.7529
Gaussian NB	0.6803	0.6803	0.6885	0.7647	0.7529
Neuro network	0.6939	0.7075	0.6230	0.7647	0.7765
Mean	0.6748	0.6857	0.6787	0.7588	0.7588

Figure 11. Crude oil price predict gasoline price, detailed classification results.

Regression models	Basic	with Additional	Period 1	Period 2 no extra	Period 2 with extra
Multi Linear	0.2873	0.2876	0.3610	0.4625	0.4741
Random Forest	0.1284	0.2695	0.1018	0.4352	0.4527
Gradient Boosting	0.2564	0.1967	-0.0615	0.4665	0.4715
XGBoost	0.0871	0.1920	-0.0462	0.3870	0.4176
AdaBoost	0.1667	0.1371	0.3466	0.3033	0.2814
Mean	0.1852	0.2166	0.1403	0.4109	0.4195

Figure 12. Crude oil price predict gasoline price, detailed regression results.

Classification models	Basic	with Additional	Period 1	Period 2 no extra	Period 2 with extra
Random Forest	0.5374	0.5918	0.6885	0.6353	0.7059
Logistic Regression	0.4490	0.6190	0.5246	0.7176	0.6941
Gradient Boosting	0.4966	0.5374	0.6557	0.5765	0.5529
XGBoost	0.5306	0.5782	0.6721	0.5765	0.4941
AdaBoost	0.4966	0.5306	0.6066	0.6118	0.6353
SVC Linear	0.5238	0.6054	0.5738	0.6941	0.7294
SVC Poly	0.5306	0.5986	0.5082	0.6941	0.6706
SVC RBF	0.4694	0.6259	0.5902	0.7176	0.6588
Gaussian NB	0.4966	0.5714	0.6885	0.6824	0.7294
Neuro network	0.5578	0.6122	0.5246	0.5765	0.6824
Mean	0.5088	0.5871	0.6033	0.6482	0.6553

Figure 13. Crude oil price predict carbon dioxide emission, detailed classification results.

Regression models	Basic	with Additional	Period 1	Period 2 no extra	Period 2 with extra
Multi Linear	-0.0210	0.0251	0.0477	0.0733	0.0627
Random Forest	-0.1339	-0.0683	0.1592	-0.0017	0.0434
Gradient Boosting	-0.1171	-0.1850	0.0833	-0.1566	-0.1266
XGBoost	-0.1832	-0.1839	0.0807	-0.2370	-0.2264
AdaBoost	-0.0156	0.0169	0.0393	0.0237	0.0139
Mean	-0.0941	-0.0790	0.0820	-0.0596	-0.0466

Figure 14. Crude oil price predict carbon dioxide emission, detailed regression results.