Does Crude Oil Markets Obtain Unrealized Predictive Capabilities?

——from Gasoline to CO_2 Concentration

Executive Summary

In 1984, Richard Roll published the paper *"Orange Juice and Weather,"* where he determined that there is a "statistically significant relation between Orange Juice returns and subsequent errors in temperature forecasts." This insight led to the conclusion that commodity markets contain valuable information in addition to the weather forecast, which shows that financial instruments are beneficial in aiding the prediction of prices of everyday goods that the general public are typically invested in.[1]

Inspired by Roll, this paper investigates three powerful insights buried within the pricing of crude oil instruments. This paper explores the statistical relationships that crude oil spot and futures prices have with gasoline, something that the general public really cares about. Based on these relationships, this paper further explores crude oil prices to see if they're capable of providing the insights for something that the general public should care about - Carbon Dioxide Concentrations (CO_2).[2]

This paper establishes meaningful connections among crude oil, gasoline, and CO_2. In addition, this paper also tests their statistical significance and uses these relationships to predict their future values through various time series models and machine learning techniques. In doing so, the three hidden insights became more evident. The first is that futures can provide evidence of statistically significant relationships between the listed prices and some other thing that impacts the general public. Second, the time frame in which these relationships are realized at the spot prices becomes evident, allowing people to grasp when these expected changes occur. Lastly, investors can take advantage of this hidden knowledge to construct portfolios that profit from these unrealized expectations.

1. Data Description

Table 1: Data Description

Variables	Frequency	Source	Period
Oil WTI	W	US EIA	Jan-1986 — Jan-2021
Oil Brent	W	US EIA	June-1988 — Jan-2021
Oil Futures	W	US EIA	Jan-1986 — Jan-2021
Gasoline Price	W	US EIA	Jan-1991 — Jan-2021
CO_2 Concentration	W	NOAA	May-1974 — Jan-2021
	M	NOAA	May-1974 — Jan-2021
	Q	NOAA	May-1974 — Jan-2021
GDP-USA	Q	OECD	Jan-1987 — Jan-2021
GDP-Europe	Q	OECD	Jan-1987 — Jan-2021
GDP-China	Q	OECD	Jan-1987 — Jan-2021
Global Temperature	M	NOAA	Jan-1958 — Dec-2020
Sugar Price	W	Macrotrends	Nov-1962 — Jan-2021
Ethanol Price	W	Investing.com	Apr-2005 — Jan-2021

([Source][3, 4] US EIA: US Energy Information Administration ; NOAA: National Oceanic and Atmospheric Administration ; OECD: Oceanic and Atmospheric Administration web-site, based on the Mauna Loa Observatory Station; MT: Macro trends; [Frequency]W: Weekly; M: Monthly; Q: Quarterly)*

2. Exploratory analysis

By generating different scenarios for each of the research questions, it becomes easier to find which time frame is the best for statistical modeling. As mentioned before, the research is conducted on three different time frames, corresponding to three different investment behaviors:

- Weekly Time Frame (Swing Investment)
- Monthly Time Frame (Value Investment)
- Quarterly Time Frame (Economic Investment)

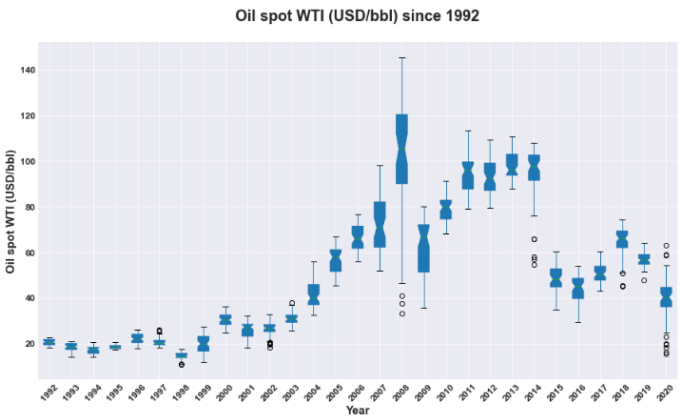
2.1 Historical Performance of Crude Oil, Oil Futures, and CO_2 Concentration

The expectation that crude oil prices directly affect gasoline prices stems from the fact that crude oil

costs are the main component of retail gasoline prices. From 2010-2020, approximately 59% of the average gasoline price was associated with the cost of crude oil. The remaining components of gasoline's retail prices were refining fees, distribution, marketing, and federal/state taxes. From a composition perspective, crude oil price fluctuations will have the most considerable impact on gasoline prices.

It is crucial to understand the crude oil supply and demand chain's primary disputers in the United States. The price of crude oil is driven by the amount of supply and demand present in the economy. This supply and demand chain has been and, for the foreseeable future, always exposed under the threat of disruption. Events, such as geopolitical and weather-related developments, play a significant role in physical turmoil and the uncertainty of possible trouble. Increases in uncertainty can lead to higher volatility in prices. Futures contracts help mitigate this uncertainty for market players by locking in the price that a buyer and seller agree upon at a future delivery date. This agreed transaction provides direct information about the market's expectation of oil futures' price and potential expectations of oil-based products such as gasoline.[5, 6]

Figure 1: Oil Spot WTI (USD/bbl) since 1992



Demand-supply and extreme market events drive WTI. Its price levels reach the highest yearly variance levels during crises, from a high correlation with financial markets. A structural break in 2008 is included in the analysis due to the high volatility of crude oil prices. Hence, the research focuses on Crude oil price derived from WTI (Figure 1).

CO_2 Concentration shows seasonal trends, with a peak in Spring and low in Autumn every year (Figure 2). CO_2 Concentration has increased every year. Different time frames: weekly, quarterly, and monthly can explain lags in time series differently. The wider the time frame, the more seasonal lag effect is aggregated, resulting in a less seasonal effect in the time series. [7]

Figure 2: Seasonal Carbon Dioxide Concentration Plot

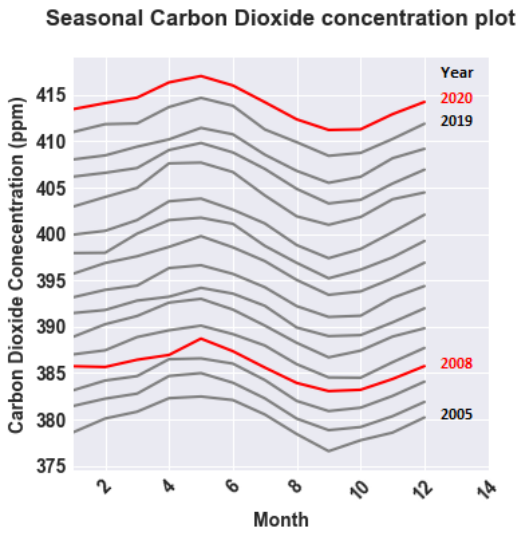


Figure 3: Augmented Dickey Fuller Test(Test statistics)

Augmented Dickey Fuller Test (t-score)

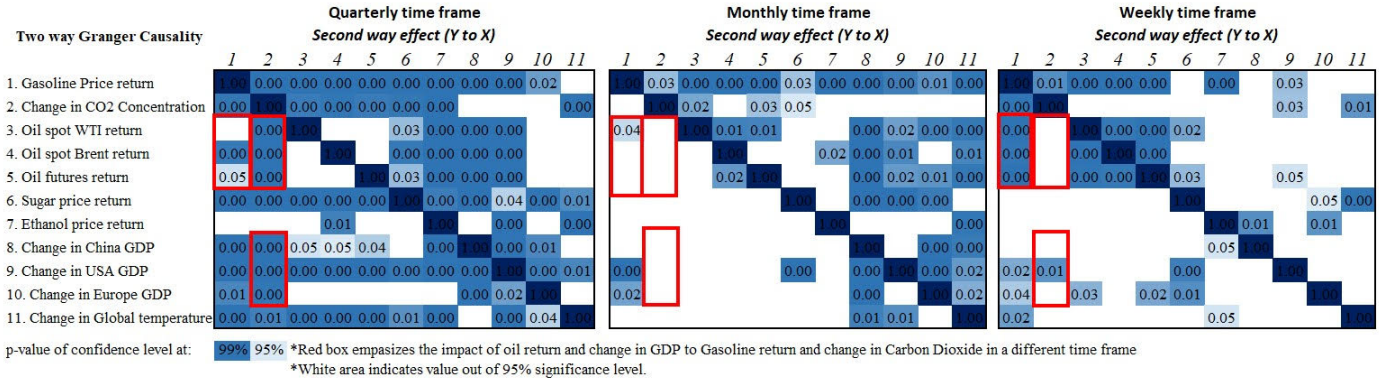
	Weekly		*Monthly*			*Quarterly*		
Variables	*Level*	*First diff*	*Level*	*First dif*	*Seasonal dif*	*Level*	*First dif*	*Seasonal dif*
Gasoline price	-1.94	-9.66***	-1.43	-6.09***	-9.00***	-1.42	-11.05***	-4.71***
CO2%	0.24	-0.15***	3.09	-4.97***	-9.58***	2.83	-2.71*	-6.45***
WTI spot	-2.37	-12.10***	-1.70	-14.06***	-6.9***	-1.79	-9.79***	-5.23***
BRENT spot	-2.12	-25.29***	-2.14	-17.79***	-8.27***	-2.21	-9.52***	-4.84***
Oil futures	-2.36	-10.91***	-1.69	-17.27***	-6.94***	-1.76	-9.53***	-5.25***
Sugar price	-2.57	-21.91***	-2.00	-15.96***	-9.57***	-2.06	-10.58***	-6.87***
Ethanol price	-3.26**	-27.80***	-3.13	-13.96***	-6.45***	-3.64	-7.48***	-4.27***
Global temp	-2.57*	-5.87***	-1.45	-9.90***	-11.4***	-1.45	-6.82***	-11.4***
USA GDP	1.46	-5.57***	0.99	-3.39**	-6.18***	1.50	-4.06***	-2.87**
China GDP	1.73	-14.62***	2.07	-2.89**	-6.36***	1.00	-2.96**	-5.22***
Europe GDP	-0.36	-5.52***	-0.37	-3.65***	-7.06***	-0.37	-5.71***	-5.05***

*Statistical significance at the 10% level **Statistical significance at the 5% level***Statistical significance at the 1% level

2.2 Stationarity

The stationarity study of variables was performed using the traditional unit root test - Augmented Dickey-Fuller test. Weekly, monthly, and quarterly time frames were tested, and the results are summarized in Figure 3. These results show that at any time frame, crude Oil (WTI) and gasoline are non-stationary *(based on the level basis t-statistic and confidence level)* and stationary in the first logarithmic difference. After taking the first logarithmic difference in the monthly and quarterly time frame for CO_2, a residual seasonal effect was found.

Figure 4: Two-way Granger Causality Test- 10 lags showing p-values



The project remedies the seasonal effect of CO_2 by taking 12-month-difference for monthly time frame and 4-quarter-difference for quarterly time frame and logarithmic difference is applied in the rest of the analysis.[8]

2.3 Granger Causality

A two-way Granger Causality test helps understand how significantly two variables explain cause and effect on each other, especially when one comes before the other in the time series *(under lagged conditions)*. The null hypothesis states that the lagged independent variable does not explain the dependent variable's variation *(does not Granger-Cause)*.

Gasoline: From Figure 4, oil spot WTI, oil spot Brent and oil futures returns impact gasoline price returns in all of the time-frames indicated by small p-values. The bi-directional relationship of gasoline returns and crude oil returns (spot and future) exists, so the Vector Auto-Regression (oil as endogenous variable) and Vector Error Correction Mechanism (oil as endogenous variable) are appropriate. In addition , the bi-directional relationship exists among gasoline, ethanol and sugar returns, suggesting that these variables should be treated as endogenous variables in Vector Auto-Regression or Vector Error correction Mechanism model *(treating oil as exogenous variable)*.

CO_2 Concentration: From Figure 4, along with oil spot WTI, oil spot Brent and oil futures returns, change in GDP also impacts the change in CO_2 Concentration in the quarterly time frame. Hence, along with the above variables, one should consider the impact of global temperature on the change in CO_2 Concentration because of its Granger causal significance suggesting that Vector based models (VAR, VECM) are appropriate.

2.4 Asymmetry of Returns and Lags

Ten lags have been picked to explore crude oil and oil futures' effect to explain gasoline price fluctuation and CO_2 Concentration. Asymmetry in price returns of crude oil was observed, with negative returns dominating over positive returns and significantly impacting gasoline returns *(with a positive correlation)*.

Gasoline and Future Gasoline: With 99% confidence, one can see that crude oil predicts gasoline's price return, up to **two lags** with a positive relationships in the weekly time frame *(For example, from Figure 5, t-scores of 4.2 and 3.3 in lag 1 and lag 2, respectively)*.

Similarly, oil futures impact future gasoline's price return, up to **two lags** with a positive correlation in the weekly time frame.

When we aggregate the time frames to monthly and quarterly levels, the prediction is significant only up to the first lag (for gasoline), further diminishing to zero lag in the quarterly time frame (for future gasoline).

CO_2 Concentration: Contrasting to gasoline's observations, crude oil predicts CO_2 Concentration significantly only with data encompassing negative returns *(asymmetry explained already)*, and the lags also expand with the time frames **(Lag 6 for Monthly and Lag 2 for Quarterly time frame)**.

2.5 Correlation

Gasoline: From the correlation plot, Figure 6, one can deduce that crude oil WTI returns, oil futures returns, and gasoline price returns are strongly cor-

Figure 5: T-scores for asymmetric multiple linear regressions, 10 lags at different time frames.

	Timeframe	Adj R-squred	Positive return: delta P+	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10	Negative return: delta P-	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9	Lag 10
Gasoline	*Weekly*	**0.19**	0.17	4.22	3.32	0.96	2.16	0.63	1.50	-0.20	1.47	1.03	0.37	3.93	5.25	5.76	2.13	0.11	-1.00	-2.02	-1.54	-0.13	0.32	0.62
	Monthly	**0.39**	6.11	4.41	0.48	0.18	0.02	-1.22	-2.84	0.50	1.22	-0.70	0.42	6.92	5.16	1.87	0.95	-1.00	-0.27	0.72	-0.37	-2.03	0.32	-0.24
	Quarterly	**0.50**	3.35	0.01	-0.91	-0.10	2.59	-0.42	-1.16	0.32	-0.18	1.19	-1.05	7.81	1.23	-0.69	-0.36	0.54	-1.28	1.33	-0.27	-0.35	0.04	-1.20
Future Gasoline	*Weekly*	**0.19**	4.08	4.00	1.26	0.86	0.74	0.89	-0.55	0.77	-0.07	-0.64	1.48	6.76	5.99	2.33	0.56	-0.31	-1.76	-1.22	-0.43	0.48	1.15	-2.79
	Monthly	**0.20**	3.08	1.27	0.00	-0.26	-0.34	-2.01	-0.20	0.61	-1.20	-0.19	2.96	5.86	-0.90	0.27	-0.48	-1.43	0.67	0.56	-2.03	0.21	-0.35	1.32
	Quarterly	**-0.10**	-0.09	-0.35	0.61	0.24	-1.26	-0.76	0.49	0.47	0.24	-1.39	0.06	-0.18	-1.43	-0.32	0.65	0.06	-0.59	-0.32	-1.26	0.11	-0.33	-0.30
CO2 Concentration	*Weekly*	**0.01**	0.12	0.03	0.97	0.12	-1.16	0.25	-0.17	-0.13	0.55	-0.59	1.93	2.11	1.03	1.52	0.50	0.30	-0.07	0.36	0.10	0.00	0.04	-0.79
	Monthly	**0.04**	0.49	0.08	-0.64	0.96	-1.55	-0.62	2.12	-0.38	-0.30	-0.52	0.70	-1.20	0.50	-0.78	1.75	-1.86	1.96	-2.66	0.95	0.37	0.34	0.73
	Quarterly	**0.09**	-0.58	-1.01	-0.83	1.28	0.59	0.69	0.45	-1.15	0.40	-0.62	0.54	-0.10	1.88	-2.84	0.51	-1.10	-0.57	0.12	0.89	0.68	-0.20	-0.13

99% confidence level 95% confidence level

Crude oil spot return (Δ P) as a predictor of gasoline return and CO_2 Concentration. oil futures return (ΔP) as a predictor of future gasoline.

Figure 6: Correlation plots of each feature with respect to Gasoline

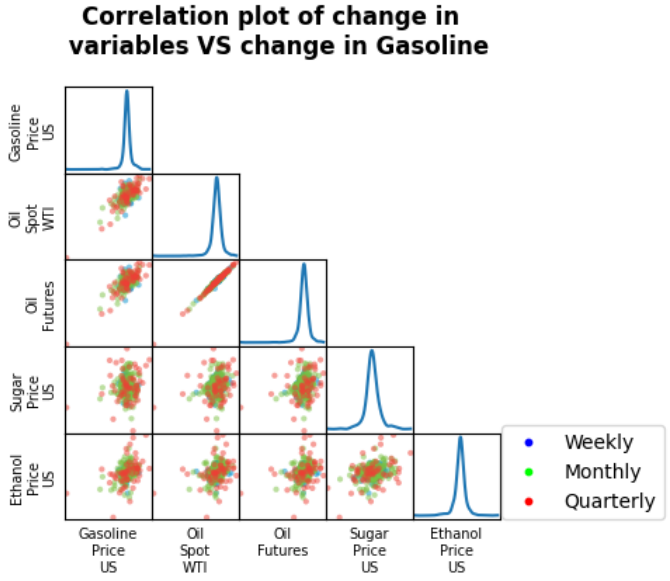
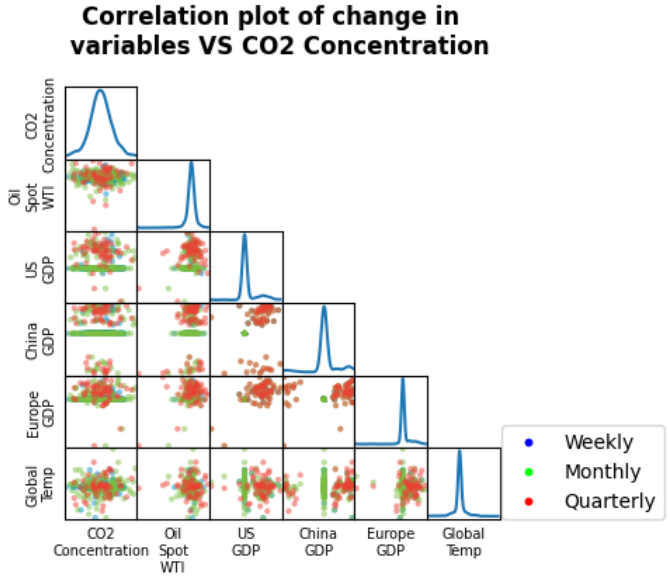


Figure 7: Correlation plots of each feature with respect to CO_2 Concentration



related. Oil futures and oil spot WTI show multicollinearity and should not be considered as independent variables together. The other variables' returns (Sugar and Ethanol) are not strongly correlated with gasoline price returns.

CO_2 Concentration: From the correlation plot, Figure 7, one can see that oil spot WTI price returns are correlated with change in CO_2 Concentration. Global Temperature and GDP show a strong correlation with the change in CO_2 Concentration in the quarterly time frame *(thus conforming to Granger-Causal results)*. Further statistical analysis is required in the quarterly time frame. The results do show a moderate relationship in all time frames.

3. Methodologies

This paper focuses on two major methods to address the research objectives: the inference method and the prediction method. The Inference method is used to find the statistical causation according to the econometric and time series modeling by investigating the sign, relationship, and magnitude of the coefficients of oil variables within models, to determine whether the oil variable has a statistical influence on

the objective variable or not.

The prediction method is used to find the prediction performance on out of sample data using time series and machine learning models. To see the performance of models with oil as a feature, models with oil features are compared with models without oil.

To strengthen the final conclusion, this research is not based solely on one model. Instead, three models for inference and five models for prediction analysis have been studied. The inference part is based on three different econometric and time series modeling: Auto-Regressive Mean with GARCH Volatility, Vector Auto-Regression(VAR), and Vector Error Correction Mechanism(VECM). The prediction part is based on five different time series and machine learning models: Linear Regression, Auto-Regressive Mean with GARCH Volatility, Vector Auto-Regression(VAR), Decision Tree Regression, and Random Forest Regression.

3.1 Inferential Analysis Models

The inferential analysis is constructed by exploring beta coefficients in direction, lag, and significance of oil related variables (oil spot and oil futures returns) across all models at different time frames. These models are:

3.1.1 Auto-regressive Time Series with GARCH Volatility (ARX-GARCH)

The Uni-variate time series is a simple model for the mean modeling. Adding Auto-regressive terms provide better-fitted results. The variance modeling is represented by the GARCH effect, which helps capture the volatility clustering.

The Mean Equation is defined as:

$$\begin{aligned} Y_t =& a + a_i \sum_{i=1}^{p} y_{t-i} + b_+ \sum_{i=0}^{nx} (X_{t-i}^{+}) + \\ & b_- \sum_{i=0}^{nx} (X_{t-i}^{-}) + c_i \sum_{i=1}^{n} z_{i,t} + \varepsilon t \end{aligned} \tag{1}$$

where Y_t is the percentage change in objective variable, such as ΔGasoline Price, ΔFuture Gasoline Price (at period $t+1$), and ΔCO_2 Concentration at time t. X_{t-i}^{+} is the percentage change in oil variable when the percentage change is positive, such as ΔOil Spot Price or ΔOil Future Price at time t. X_{t-i}^{-} is the percentage change in oil variable when the percentage change is negative, such as ΔOil Spot Price, and ΔOil Future Price at time t. $\sum_{i=1}^{n} z_{i,t}$ represents other features that will be included into the model, which includes financial crisis, other future prices, global temperature, and GDP by countries.

The Variance Equation is defined as:

$$\begin{aligned} \sigma^2 t =& \omega + \alpha \sum_{i=1}^{q} \varepsilon_{t-q}^2 + \\ & \beta \sum_{i=1}^{p} \sigma_{t-p}^2 + \zeta t \end{aligned} \tag{2}$$

where σ is the standard deviation of the percentage change in objective variable, such as ΔGasoline Price, ΔFuture Gasoline Price (at period $t+1$), and ΔCO_2 Concentration at time t. $\sum_{i=1}^{q} \varepsilon_{t-q}^2$ represents the lag of error terms ε^2 from time t to time t-q. $\sum_{i=1}^{p} \sigma_{t-p}^2$ are terms represent the σ^2 from time t to time t-p.

3.1.2 Vector Auto-Regression (VAR-EN,VAR-EX)

The research has shown from Granger causality that the bi-causation relationship between objective variables and oil variables exists with the variable dynamics. The model that best suits the current situation is Vector Auto-Regression treating oil as the endogenous variable (VAR-EN). However, the paper from Lucio Carpio [9] suggested that the variable, oil price percentage change, suits as the exogenous. In this case, the endogenous variable contains other features, such as the ethanol and sugar price change, representing substitution and cost-effectiveness. Therefore we developed Vector Auto-Regression and treated the oil variable as the exogenous variable (VAR-EX).

The vector equation for the VAR model is defined as:

$$Y_t = A \sum_{i=i}^{p} Y_{t-i} + B \sum_{i=i}^{n} Z_{t,i} + Et \tag{3}$$

Where Y_t represent vector of endogenous variable. $\sum_{i=i}^{p} Y_{t-i}$ are the lag of vector of endogenous variables from time t-1 to time t-p. $\sum_{i=i}^{n} Z_{t,i}$ are the vector of exogenous variables. E_t is the vector of error terms.

3.1.3 Vector Error Correction Mechanism (VECM-EN)

In addition to the multivariate causation of variables, the relationship of oil and objective variables

might have co-integrating relationships. The best-suited model for this situation is the Vector Error Correction Mechanism. The Error Correction Mechanism equation indicates the short-term effects.[10]

The VECM equation is defined as:

$$\Delta y_t = \beta_0 + \sum_{i=1}^{p} \beta_i \Delta y_{t-i} + \sum_{i=0}^{n} \delta_i X_{t-i} + \phi z_{t-1} + \mu_t \tag{4}$$

Where Y_t is the vector of endogenous variable. $\sum_{i=1}^{p} \Delta y_{t-i}$ are the lag of vector of endogenous variables from time $t-1$ to time $t-p$. $\sum_{i=0}^{n} X_{t-i}$ are the vector of exogenous variables.

3.2 Prediction Models

The main objective in prediction is to compare the model performance with or without oil variable as predictors. The research constructs five different models to predict the objective variable in unseen data. They are Linear Regression (LR), Auto-Regressive Mean with Garch Volatility (ARX-GARCH), Vector Auto-Regression (VAR), Decision Tree Regression (DT) and Random Forest Regression (RF). Most of these models are constructed based on four different scenarios *(More explanation in section 4)*. The models use training data set for training the model (equation 5) and evaluate them on two different testing data sets by RMSE metric (see equations 6 , 7).

$$F(x) = f(x_{train}) \tag{5}$$

$$RMSE_1 = \sqrt{\sum_{i=1}^{n} \frac{(F(x_{test1,i}) - y_{test1,i})^2}{n}} \tag{6}$$

$$RMSE_2 = \sqrt{\sum_{i=1}^{n} \frac{(F(x_{test2,i}) - y_{test2,i})^2}{n}} \tag{7}$$

4. Statistical Analysis

4.1 Inference Analysis

For statistical inference modeling, the ARX-GARCH, VAR (Both VAR models), and VECM (oil as endogenous) were used to develop the statistical significance for the relationships between spot prices of crude oil and gasoline, future prices of crude oil and gasoline, crude oil spot price and CO_2 Concentrations respectively. All these variables were first transformed into returns. These models were fit based on the weekly, monthly, and quarterly time frame intervals to develop insights and analyze whose lags can produce significant relationships. ARX-GARCH and VAR (oil as exogenous) models were also created to reflect how the relationships behaved when returns were increasing and decreasing *(indicated through the columns $\beta+$ and beta$-$)*. The results of these models are summarized in Table 2, Table 3, and Table 4 for each time frame.

Table 2: Weekly Inference Analysis Result

Objective-Y / Explanatory-X		Gasoline Oil Spot		Future Gasoline Oil Futures		CO_2 Oil Spot	
Weekly	Lag	$\beta+$	$\beta-$	$\beta+$	$\beta-$	$\beta+$	$\beta-$
ARX-GARCH	0	-0.050*	0.064*	0.269***	0.142***	-0.92**	
	1	0.253***	0.138***	0.088**	0.124***	0.03	
	2	0.066**	0.110***	0.050	0.040	-0.55*	
VAR-EN	1	0.179***		0.181***		-0.35	
	2	0.135***		0.138***		-0.03	
VAR-EX	0	0.0004	0.0114	0.237***	0.144***	-0.32	-0.05*
	1	0.217***	0.118***	0.141***	0.158***	-0.60	-0.35
	2	0.123***	0.151***	0.040**	0.052*	-0.19	-0.12
	3	0.028	0.058**	0.066	0.017	0.58	-0.48
	4	0.050*	0.019	0.004	0.004	0.34	-0.62
VECM-EN	1	0.161***		0.154***		No Cointegration	
	2	0.113***		0.120***			

*Statistical significance at the 10% level **Statistical significance at the 5% level***Statistical significance at the 1% level

Table 3: Monthly Inference Analysis Result

Objective-Y / Explanatory-X		Gasoline Oil Spot		Future Gasoline Oil Futures		CO_2 Oil Spot	
Monthly	Lag	$\beta+$	$\beta-$	$\beta+$	$\beta-$	$\beta+$	$\beta-$
ARX-GARCH	0	0.469***	0.431***	0.574***	0.584***	-0.00	
	1	0.301***	0.144	0.056	0.043	0.00	
	2	0.058	0.186*	0.202*	0.166*	-0.00	
	3					0.00	
	12					-0.00***	
VAR-EN	1	0.220***		0.257***		0.00	
VAR-EX	0	0.395***	0.547***	0.513***	0.684***	0.00	-0.00
	1	0.262**	0.157*	0.144	0.053	0.00	0.00
	2	0.108	-0.073	0.139	-0.132*	0.00	0.00
	3					0.05	-0.00
VECM-EN	1	0.104		0.140*		No Cointegration	

*Statistical significance at the 10% level **Statistical significance at the 5% level***Statistical significance at the 1% level

Table 4: Quarterly Inference Analysis Result

Objective-Y / Explanatory-X		Gasoline Oil Spot		Future Gasoline Oil Futures		CO_2 Oil Spot	
Quarterly	Lag	$\beta+$	$\beta-$	$\beta+$	$\beta-$	$\beta+$	$\beta-$
ARX-GARCH	0	0.546***	0.549***	0.611***	0.623***	0.00	
	1	0.130	0.180*	0.161	-0.008	0.00	
	8					0.00***	
	11					-0.00*	
VAR-En	1	0.409**		0.354		0.00	
VAR-EX	0	0.487***	0.790***	0.477***	0.833***		
	1	0.133	-0.099	0.090	-0.119		
	2	0.055	-0.016	0.209	-0.032		
	4	0.054	0.100	0.079	0.112	-0.00	-0.00*
	6					0.00	-0.00**
VECM-EN	1	0.365		0.385		No	
	2	-0.268		-0.097		Cointegration	

*Statistical significance at the 10% level **Statistical significance at the 5% level***Statistical significance at the 1% level

Models showing significant coefficients *(with 99% confidence levels)* are presented in the first and second lag for the weekly time frame relationship between crude oil and gasoline returns. For lags 1 and 2, significant and positive coefficients indicate that the change in crude oil spot prices will result in the same direction move in gas prices, which will be realized approximately 1-2 weeks later. The magnitude of these coefficients is also different for each asymmetric profile in the changes of prices. At lag 1, for the ARX-GARCH and VAR *(oil as exogenous)* models, it can be noted that the cost of gasoline changes corresponding to the movements of crude spot oil, which follows the expectation of the general public and the latest journals.[11, 12]

For the weekly time frame relationship between crude oil future returns and future gasoline returns, the models have significant coefficients *(with 99% confidence)* from lag 0 to lag 2. As the relationship between spot and gasoline returns becomes more evident, these coefficients are positive and have asymmetric profiles. Additionally, the models suggest that the relationship can be realized immediately. In summary, both oil spot and futures returns present a quick 0 to 2 weeks transition into gasoline returns.[13]

For the weekly time frame relationship of oil future returns and carbon dioxide returns, the models showed no significant coefficients except for the ARX-GARCH at lag 0. However, this coefficient is negative and indicates that when oil future returns increase, the CO_2 Concentration will decrease, which contributes to the instant increase in CO_2 emissions from higher production costs. The monthly and quarterly time frames were explored to retrieve more meaningful results for this relationship.

For the monthly time frame of relationships between crude oil returns (spot and futures) and gasoline prices (current and future), every model showed a significant coefficient at lag 0, but only few have statistically significant coefficients in lag 1 (Table 3). The above finding also aligns with the analysis that was done at the weekly time frame whose transmission usually take place around 0-2 weeks.

For the monthly CO_2 Concentrations, only the ARX-GARCH model showed a significant coefficient at lag 12. Although this coefficient was small and negative, it indicates that the oil futures returns may have impacted CO_2 Concentrations at a longer time frame. Therefore, the analysis of quarterly time frame is necessary.

As shown in Table 4, the relationship between crude oil (spot and futures) and gasoline prices at a quarterly time frame aligns with the previous analysis. The models show significant coefficients in lag 0 once again, which further indicate that the relationship has instant effects.

For the CO_2 coefficients at the quarterly time frame, the models begin to depict significant coefficients at lags 4, 6, 8, and 11. However, these lags are almost all negative and small in magnitude, suggesting that the optimal lag predicting CO_2 is indicated in a more broad time frame. Therefore, among all three time frames, the longer horizon "quarterly" is the time frame that indicates the slow transmission between crude oil and CO_2 Concentration.

In conclusion, one can confirm that **crude oil spot and futures' expectation affect gasoline prices in a zero to two-week time frame with confidence.** This short-term effect explains why people generally expect this to happen. Almost all three time frames and both asymmetric profiles have positive significant coefficients *(in different magnitude)*.

On the other hand, the relationship between crude oil and CO_2 Concentrations is classified as negative relationships **(as oil return increase the CO_2 Concentration return will move to the opposite direction)**. It takes a much longer time for their relation-

ship to form which explains why people are not generally expected this relationship to happen.

4.2 Prediction Analysis

Table 5: Weekly Prediction Analysis Result

Weekly	Variables	Gasoline + Oil RMSE 1	Gasoline + Oil RMSE 2	Gasoline + Oil(Future) RMSE 1	Gasoline + Oil(Future) RMSE 2	Oil and CO_2 RMSE 1	Oil and CO_2 RMSE 2
Linear Regression	Only Y	0.0119	0.0167	0.0124	0.0163	0.647	0.581
	Y and Oil	0.0100	0.0127	0.0103	0.0177	0.645	0.582
	All Features w/o Oil	0.0121	0.0207	0.0129	0.0196	0.643	0578
	All Features	0.0101	0.0126	0.0104	0.0219	0.641	0.579
SARIMA with GARCH Volatility	Only Y	0.0117	0.0202	0.0117	0.0202	0.694	0.651
	Y and Oil	0.0103	0.0234	0.0098	0.0188	0.696	0.653
	All Features w/o Oil	0.0133	0.0182	0.0143	0.0172	0.689	0.645
	All Features	0.0105	0.0246	0.0094	0.0185	0.691	0.647
VAR	Oil as Endogenous	0.0103	0.0239	0.0101	0.0192	0.686	0.635
	Oil as Exogenous	0.0098	0.0170	0.0105	0.0166	0.698	0.641
	No Oil	0.0125	0.0273	0.0125	0.0170	0.701	0.640
Decision Tree	Only Y	0.0134	0.0151	0.0165	0.0139	0.727	0.661
	Y and Oil	0.0132	0.0214	0.0181	0.0212	0.853	0.763
	All Features w/o Oil	0.0133	0.0104	0.0144	0.0231	0.731	0.721
	All Features	0.0147	0.0147	0.0145	0.0188	0.852	0.718
Random Forest	Only Y	0.0126	0.0147	0.0134	0.0151	0.639	0.573
	Y and Oil	0.0115	0.0153	0.0147	0.0241	0.653	0.586
	All Features w/o Oil	0.0123	0.0152	0.0145	0.0188	0.644	0.586
	All Features	0.0153	N/A	0.0149	0.0216	0.652	0.589
		Best	Worst	Best	Worst	Best	Worst
Summary	With Oil	7	4	6	5	4	7
	No Oil	3	6	4	5	6	3
	Scaling Ratio	2.33	0.67	1.50	1.00	0.67	2.33

This research used five models that are described in the methodologies section. The models were run at three different time frames (weekly, monthly, and quarterly) and used three to four different sets of features as agent to predict the future gasoline price returns and CO_2 Concentration returns. These data sets were split into three sets (Training, Testing set1, Testing set2). Two of the test sets were used for the generalized result, but only one test set was used for the quarterly time frame due to limitation of data spectrum. The prediction performance was evaluated as RMSE (Root mean square error). The Lower RMSE in test sets the better the model. Four of the five models were conducted using four different feature sets:

- "Only Y": the lag of objective variable
- "Y and Oil": the lag of objective variable + the oil variable lag of them
- "Every Feature no Oil": the additional explanatory variable for each objective
- "Every feature": all of the variables will be included as the features

The VAR (Vector Auto-Regression) model used three different self-explanatory feature sets, "Oil as Endogenous," "Oil as Exogenous," and "No Oil."

Table 6: Monthly Prediction Analysis Result

Monthly	Variables	Gasoline + Oil RMSE 1	Gasoline + Oil RMSE 2	Gasoline + Oil(Future) RMSE 1	Gasoline + Oil(Future) RMSE 2	Oil and CO_2 RMSE 1	Oil and CO_2 RMSE 2
Linear Regression	Only Y	0.052	0.070	0.015	0.023	0.00181	0.00185
	Y and Oil	0.045	0.111	0.018	0.046	0.00197	0.00182
	All Features w/o Oil	0.054	0.056	0.016	0.022	0.00189	0.00177
	All Features	0.043	0.107	0.048	0.111	0.00205	0.00176
SARIMA with GARCH Volatility	Only Y	0.067	0.079	0.067	0.079	0.00455	0.00379
	Y and Oil	0.043	0.111	0.056	0.111	0.00441	0.00365
	All Features w/o Oil	0.067	0.068	0.067	0.062	0.00397	0.00368
	All Features	0.041	0.109	0.054	0.108	0.00433	0.00366
VAR	Oil as Endogenous	0.041	0.082	0.047	0.051	0.00252	0.00213
	Oil as Exogenous	0.040	0.121	0.053	0.114	0.00257	0.00211
	No Oil	0.057	0.076	0.057	0.076	0.00249	0.00212
Decision Tree	Only Y	0.052	0.091	0.034	0.044	0.00302	0.00297
	Y and Oil	0.061	0.078	0.039	0.024	0.00309	0.00288
	All Features w/o Oil	0.109	0.075	0.042	0.042	0.00304	0.00288
	All Features	0.057	0.083	0.071	0.068	0.00352	0.00302
Random Forest	Only Y	0.052	0.072	0.027	0.028	0.00226	0.00203
	Y and Oil	0.045	0.057	0.025	0.022	0.00232	0.00215
	All Features w/o Oil	0.050	0.067	0.027	0.028	0.00219	0.00202
	All Features	0.047	0.055	0.046	0.038	0.00227	0.00202
		Best	Worst	Best	Worst	Best	Worst
Summary	With Oil	5	3	6	8	5	7
	No Oil	5	7	4	2	5	3
	Scaling Ratio	1.00	0.43	1.50	4.00	1	2.33

The results from models at the weekly, monthly, and quarterly time frame are summarized in Table 5, Table 6, and Table 7. These results were then further consolidated into performance ratios at the bottom of each table. These performance ratios were created by counting the models that performed the best when using features with oil versus models that performed the best without oil features. The same was done for the worst performing models with or without oil features. A performance ratio more than value 1 under the "Best" column indicates that more models with the oil features performed better than those without oil features, and vice versa. To conclude, oil and oil futures are capable of predicting gasoline prices and CO_2 Concentrations. The "Best" column ratios should be greater than 1, and the ratios for the "Worst" column should be less than 1.

With this being said, it can be seen that the weekly time frame for predicting gasoline prices with crude oil spot prices is as expected as its performance ratios of 2.33 for the "Best" and 0.66 for the "Worst". Oil futures seem to contain information to predict future gasoline prices at a weekly time frame, but because of the uncertainty of the future gasoline variable and the volatility of the futures price, the predictor contains noise which can eventually lead to over-fitting models. As for CO_2 Concentrations, it can be seen that oil futures do not include additional information to predict at the weekly time frame. The prediction performance is inconclusive and overwhelmed with

Table 7: Quarterly Prediction Analysis Result

Quarterly	Variables	Gasoline + Oil RMSE	Gasoline + Oil(Future) RMSE	Oil and CO_2 RMSE
Linear Regression	Only Y	0.100	0.037	0.0008
	Y and Oil	0.184	0.049	0.0017
	All Features w/o Oil	0.092	0.038	0.0033
	All Features	0.188	0.205	0.0031
SARIMA with GARCH Volatility	Only Y	0.090	0.090	1.1165
	Y and Oil	0.258	0.209	0.0023
	All Features w/o Oil	0.068	0.100	0.0032
	All Features	0.272	0.183	0.0039
VAR	Oil as Endogenous	0.242	0.115	0.00141
	Oil as Exogenous	0.398	0.300	0.00148
	No Oil	0.033	0.033	0.00154
Decision Tree	Only Y	0.146	0.058	0.0019
	Y and Oil	0.121	0.060	0.0017
	All Features w/o Oil	0.177	0.062	0.0030
	All Features	0.086	0.188	0.0028
Random Forest	Only Y	0.129	0.055	0.00150
	Y and Oil	0.056	0.030	0.00152
	All Features w/o Oil	0.094	0.050	0.00154
	All Features	0.048	0.064	0.00147
Summary		Best/Worst	Best/Worst	Best/Worst
	With Oil	2/3	1/4	4/1
	No Oil	3/2	4/1	1/4
	Scaling Ratio	0.67/1.50	0.25/4	4/0.25

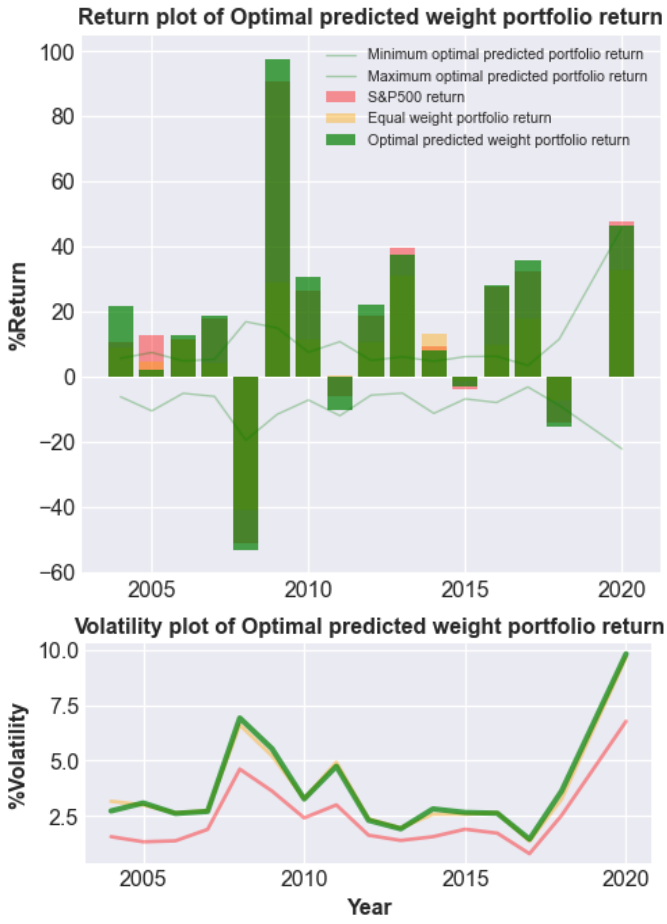
statistical noise.

As the time frames increase to monthly and quarterly, the performance metrics for predicting gasoline from crude oil spot or futures prices begin to degrade. The Best and the Worst ratios decrease and increase corresponding to the time frame. However, larger time frames benefit the performance metrics on oil spot predicting CO_2 Concentrations. **This leads to the conclusion that oil and oil futures returns are good predictors of gasoline prices and gasoline price returns at weekly basis. The crude oil returns has the potential to be a great predictor for CO_2 Concentrations returns in the quarterly time frame or even longer horizon.**

5. Investment Strategy

The oil price and CO_2 Concentration change has a long-term impact that the general person does not expect. Therefore, this research might leverage this effect to an environmental-friendly investment strategy with a suitable time frame. With the given information, one can construct a portfolio and reallocate assets quarterly (to match time frame). A simulated portfolio used in this research will consist of 4 stocks and an SP 500 index: 1. Trane Technologies (TT) offering energy-efficient climate-control systems, 2. Rockwell Automation (ROK) owns division that helps companies monitor energy usage and waste,

Figure 8: Return and Volatility Plot of Optimal Predicted Weight Portfolio



3. Hubbell (HUBB) does business in energy-efficient lighting and Schneider Electric, 4. ON Semiconductor (ON) sells power-management chips used in the electric-vehicle space.

First, to create the optimal portfolio, the max shape ratios portfolio will be selected from 1,000 simulated portfolios as the ideal portfolio per quarter. Second, features include oil features, CO_2 Concentration, and additional features for CO_2 analysis. Third, create machine learning models using features described to predict the optimal weight of the portfolio. Finally, evaluate the optimal predicted portfolio by comparing it to the equal-weighted portfolio and the SP 500 market index.

The performance presents a **significant margin** over other strategies by comparing their compounded annually returns, according to Figure 8. It wins 8 out of 16 years as the annual best portfolio and loses 5 out of 16 years as the worst portfolio. With this

attractive return, the portfolio's risk is not bad—the maximum standard deviation of 8 weeks equals to the equally weighted portfolio. In extreme case, the weekly maximum returns draw down is the worst 6 years out of 16 compared to the similarly weighted 10 years out of 16. However, the optimal predicted portfolio's weekly maximum upside is 12 years out of 16 in total, while the equaled weight portfolio only get 4 out of 16.

The predicted portfolio outperformed other portfolios because of the expected weight contains the future CO_2 *(which the public does not generally care about but they should)*. In conclusion, the investor can successfully leverage this strategy, "Optimal Predicted Environmental-Friendly Portfolio," even before the information becomes something that "people generally expect."

6. Conclusion

The analysis shows three reasons why the futures market is an extremely powerful tool. First, futures are capable of providing evidence of statistically significant relationships between the listed futures prices and some other things that impact the general public. This paper proved through the relationship of crude oil prices and gasoline, then apply this methodology to CO_2 Concentrations. Second, the time frame analysis allows people to grasp additional information when these expected changes are going to occur. When relationships form, the spot prices become evident. Lastly, investors can take advantage of this hidden knowledge to construct portfolios that profit from these unrealized expectations. Overall, the futures market contains powerful predictive information that can not only be used to obtain profits but also leads the general public to become more conscious of increasingly concerning topics such as climate change.

References

[1] Richard Roll. Orange juice and weather. *The American Economic Review*, 74:861–880, 1984.

[2] Oil and petroleum products explained. *U.S. Energy Information Administration*, page 10.09.2020, Accessed 27.02.2021.

[3] Gasoline and diesel fuel update. *U.S. Energy Information Administration*, page 22.02.2021, Accessed 27.02.2021.

[4] Observation package (obspack) data products. *Earth System Research Laboratories Global Monitoring Laboratory*, page 12.08.2019, Accessed 27.02.2021.

[5] Tobechi F., Chinazaekpere N., Uwazie I., Lasbrey I., and Ikwor O. Oil price, energy consumption and carbon dioxide (co2) emissions: insight into sustainability challenges in venezuela. *Latin American Economic Review*, 28:8, 2019.

[6] Hurricane harvey caused u.s. gulf coast refinery runs to drop, gasoline prices to rise. *U.S. Energy Information Administration*, page 11.09.2017, Accessed 27.02.2021.

[7] Charles D. Keeling and Timothy P. Whorf. Atmospheric carbon dioxide concentrations at 10 locations spanning latitudes 82°n to 90°s. *Scripps Institution of Oceanography*, NDP 001a, 2004.

[8] Frank Laudicinaa, Florent McIsaaca, and Marius-Cristian Frunzaa. Influence of weather variability on the orange juice prices. *Sorbonne University, Maison des Sciences Economiques*, pages 106–112, 2012.

[9] Lucio Carpio. The effects of oil price volatility on ethanol, gasoline, and sugar price forecasts. *Energy ELSEVIER*, 181:1012–1022, 2019.

[10] Rana Abdullah Ahmed and Ani Bin Shabri. Daily crude oil price forecasting model using arima, generalized autoregressive conditional heteroscedastic and support vector machines. *American Journal of Applied Sciences 11*, 3:425–432, 2014.

[11] Talel Boufateh. The environmental kuznets curve by considering asymmetric oil price shocks: evidence from the top two. *Environmental Science and Pollution Research*, 26:706–720, 2019.

[12] Andrea Bastianin, Marzio Galeotti, and Matteo Manera. Forecasting the oil–gasoline price relationship: Do asymmetries help? *Energy Economics*, 46:S44–S56, 2014.

[13] Severin Borenstein, A. Cohn Cameron, and Richard GiJbert. Do gasoline prices respond asymmetrically to crude oil price changes? *NBER WORKING PAPER SERIES*, No. 4138, 1992.