Dip a Crystal Ball into the U.S. Oil Market and Carbon Emissions

Marshall Guan[1] Ying Guan[1] Bojun Li[1] Jingxuan Li[1] Ziyue Su[1] Jialiang Wei[1] Sasha Stoikov[2]
Cornell University
{jg2262, yg532, bl755, jl4267, zs366, jw2684, sfs33}@cornell.edu

***Abstract*—In this paper, we successfully predict the gasoline prices in populated U.S. states and make robust inferences in the change of gasoline prices. We also investigate the relationship between the oil market and CO_2 concentration. In our research, we implement a state-of-the-art machine learning algorithm to accurately predict the gasoline prices in selected states within an error range of 15 cents per gallon. Moreover, we effectively implement deep learning methods and achieve a 77% accuracy in classifying the gasoline price movements. In addition, we conduct a comprehensive analysis on the factors affecting CO_2 concentration and recognize gasoline sales as the most influential factor in predicting CO_2 concentration. These results will add significant value to consumers and provide guidance for policy making.**

I. Introduction

For a very long time, people are unclear about the relationship between crude oil and gasoline prices, something that most Americans care about in their everyday lives. What does $60 per barrel mean to the oil price we need to pay at the gas station? Why is gasoline more expensive than crude oil? Is there a way to know how much the gasoline price will be tomorrow at my local pump? Or even bigger questions such as how gasoline can impact the environment and our future generations. In this paper, we will first dig into the relationship between crude oil futures, gasoline prices, and other factors. We will then explore the topic that everyone concerns: should I fill up my tank today or wait until next week? In the end, we will assess the impact of crude futures prices on the atmospheric CO_2 concentration and explore the important factors that can affect the environment.

In section II, we conduct an exploratory data analysis and present the correlations and time series of some financial data. In section III, we implement the extreme gradient boosting algorithm to predict future gasoline prices. In section IV, we deploy both logistic regression and recurrent neural networks to analyze the direction of price movement in gasoline prices. In section V, we explore the relationship of CO_2 and other important financial, seasonal indicators.

II. Financial Data Analysis

A. Data Description

We scrape the data via Bloomberg Terminal and the EIA[1] and NOAA[2] websites to carry out empirical studies on the oil market and carbon emissions. Our data sets consist of:

- Weekly crude oil futures prices of the 1-month, 2-month, 3-month, and 4-month contracts from 01/07/2000 to 02/05/2021.
- Weekly all-grade, mid-grade, regular, and premium gasoline prices at the pumps in California, Colorado, Florida, Massachusetts, Minnesota, New York, Ohio, Texas, and Washington from 01/03/2000 to 02/01/2021.
- Weekly atmospheric CO_2 concentration measured at NOAA's Mauna Loa Observatory in Hawaii from 01/07/2000 to 12/27/2019.

Weekly data is given by averaging the daily data within each week. We use weekly data universally because it reflects a reasonable time frame and mitigates the effects of sudden, volatile daily swings. Furthermore, since we notice a fair amount of missing data for dates before the year 2000, we clean the missing data and filter to only analyze the last two decades. Thus, we choose the first week of 2000 as our starting point.

B. Data Visualization

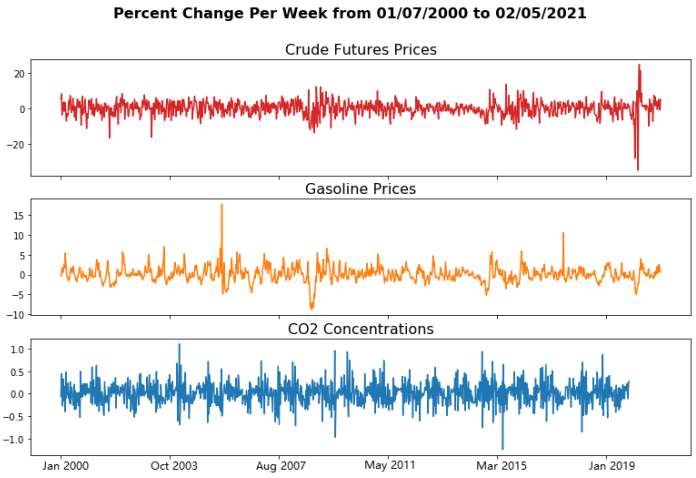

Fig. 1. Gasoline prices will move considerably after crude futures prices are highly volatile. What is the lag?

Figure 1 suggests the weekly percentage changes of crude futures prices, gasoline prices, and CO_2 concentration since 2000. Regarding percentage of change, crude futures price is most volatile, and CO_2 concentration is least volatile. Moreover, when crude futures are highly volatile, gasoline prices will also move dramatically in the future. It leads us

[1]U.S. Energy Information Administration
[2]National Oceanic and Atmospheric Administration

to explore the relationship between crude futures and gasoline prices.

It is possible that crude prices contain predictive power for future gasoline prices but not for CO_2. And what is the lag? To figure this out, we will first look into the correlations of gasoline prices by region of pumps and maturity of crude futures. We decide to analyze the all-grade gasoline rather than other grades because it provides a more comprehensive picture of the gasoline price.

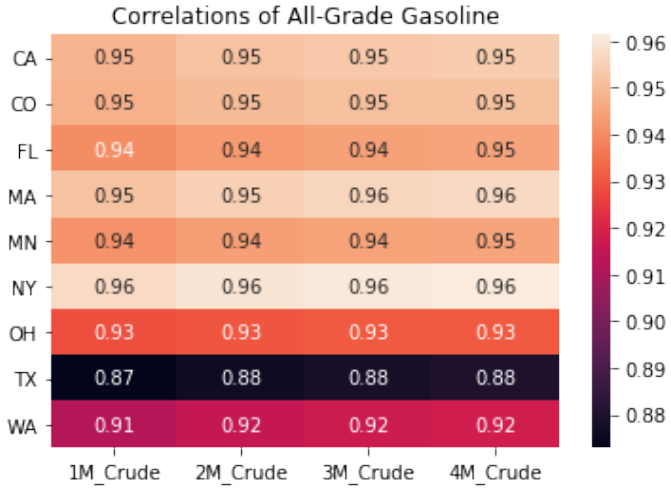

Fig. 2. Texas has the lowest correlations in general.

From Figure 2, we observe that Texas has a relatively lower correlation with crude prices than other regions. Moreover, a longer term of crude futures to maturity seems to have a slightly higher correlation.

III. Predicting Gasoline Price Levels

A. Model Methodology and Model Selection

Many of the literature reviews suggest that crude oil futures may have predictive power for future gasoline prices. The correlation analysis in section II also indicates a strong relationship between crude oil futures and future gasoline prices. Thus, we deploy several machine learning methods to examine their relationship, relative importance, and prediction performance.

We design our models with five time lags (concurrent, 1 month, 2 months, 3 months and 4 months) between the future gasoline spot prices and present crude oil futures prices. Our response variable is the future gasoline spot price in nine different states, and the X space is present crude oil futures prices with 1,2,3,4-month expiration. In total, we run 45 models (5 different lags by 9 states) for each machine learning method. For instance, the data sets with a 2-month lag match the gasoline spot price two months later with present crude oil futures prices across four different expiration dates. After exploring the data, we decide to examine it with LASSO, multivariate adaptive regression splines (MARS), and gradient boosting method (XGBoost). Through 80-20 training and test data split, we ensure that our models are not biased towards any specific periods. When training each model, we deploy 5-fold cross validation to ensure each model is robust and consistent. The mean squared prediction error (MSPE) metric shows that XGBoost outperforms LASSO and MARS by 1.1 and 15 times, respectively. Thus, we decide to use XGBoost to fit our data in predicting future gasoline prices.

XGBoost is a powerful tool to effectively predict data with great flexibility and work with non-linear numerical values such as seasonality. Our models outperform most other regression models by sequentially boosting the trees based on the max gradient of the loss function. Below is the algorithm of gradient boosting algorithm.

Algorithm 1 Gradient Boosting

1: **Initialize**

$$\mu^{(0)}(x) = \arg\ \min_\theta \sum_{i=1}^{n} L(y_i, \gamma) \tag{1}$$

2: **For** m = 1 **to** M:

a: **For** i = 1,2,...,n compute

$$r_i^{(m-1)} = -\frac{\partial L(y_i, \mu(x_i))}{\partial \mu(x_i)}\bigg|_{\mu=\mu^{(m-1)}} \tag{2}$$

b: Fit a regression tree to the targets $r_i^{(m)}$ giving terminal regions R_{jm}, j = 1,...,J.

c: For j = 1,...,J compute

$$\gamma_{jm} = \arg\ \min_\gamma \sum_{x_i \in R_{jm}} L(y_i, \mu^{m-1}(x_i) + \gamma) \tag{3}$$

d: Update $\mu^m(x) = \mu^{m-1}(x) + v \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$

3: **Output** $\hat{\mu}(x) = \mu^M(x)$

B. Model 1: Prediction using only Futures Contracts

In this model, we have the following as our X and Y inputs:

- Y_i: Future gasoline price for state i.
- X_1: 1-month expiration crude oil prices.
- X_2: 2-month expiration crude oil prices.
- X_3: 3-month expiration crude oil prices.
- X_4: 4-month expiration crude oil prices.

In Figure 3 below, the heat map represents the MSPE of our 45 models. The relative importance table shows the rank of importance for each feature in each model, with the top-ranked feature placed at the front and the feature with the least importance placed at the back. From the MSPE table, we can see that the models with a 1-month lag are the best at predicting future gasoline spot prices since they have the smallest MSPE across all states. In the relative importance table, we can see for the lag period of 3 and 4, and in most cases, the 4-month futures contract is the most influential predictor. In addition, it seems to exhibit a linear relationship between the MSPE and lag period, which proves that the model is consistent and valid. From the MSPE table, we can also observe that Texas has a relatively higher MSPE than

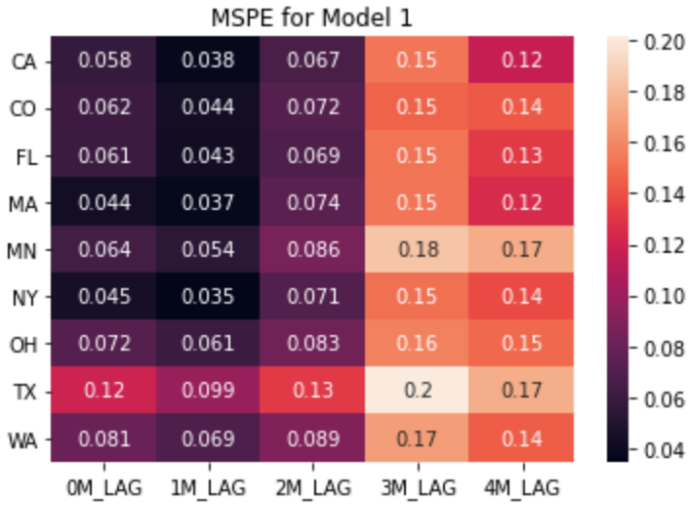

Fig. 3. 1-month-lag models tend to outperform in all states.

TABLE I
MODEL 1 RELATIVE IMPORTANCE - LONG-LAG MODELS TEND TO HAVE STABLE RELATIVE IMPORTANCE FACTORS

Region	Relative Importance				
	0M_LAG	*1M_LAG*	*2M_LAG*	*3M_LAG*	*4M_LAG*
CA	1342[a]	2413	1243	4123	4312
CO	2413	2413	1423	4123	1432
FL	1432	1432	2413	4123	2314
MA	1432	3412	2431	4312	4123
MN	1432	2413	1342	4123	4312
NY	1342	3142	4123	4123	4312
OH	3241	1324	4213	4123	4123
TX	4132	1432	4312	4213	4312
WA	1432	2413	2413	4132	4123

[*] 1,2,3,4: Time to expiry (in month) of oil futures.
[a] For model fitted with 0 month lag predicting CA gasoline prices, 1 month future prices have the most predictive power and 2 month future prices have the least predictive power.

other states. We infer that as a U.S. crude oil center, Texas amplifies the effect of factors affecting gasoline prices. We will discuss this further in model 2.

All of these findings suggest crucial economic interpretations and implications. Firstly, we expect that a lag of the closest term would have the highest level of predictability. Since gasoline and oil transactions are dominated by the futures market rather than the spot market, the model with a 1-month lag would be expected to have the highest predictability. Secondly, as the 1-month futures have the highest trading volumes compared to others, the price would be affected by many factors other than gasoline-related causes. For example, one of such factors could be people speculating on the oil market. Therefore, it is not always true that the crude oil futures with a maturity that matches the lag period are the most influential predictor in the short term. However, the long-term models tend to stick with the long-term futures, which implies that gasoline producers are more involved in the 4-month crude oil futures market.

C. Model 2: Adding three more parameters (IEO, MVCRAK and Seasonality)

In this model, we have the following as our X and Y inputs:

- Y_i: Future gasoline price for state i.
- X_1: 1-month expiration crude oil prices.
- X_2: 2-month expiration crude oil prices.
- X_3: 3-month expiration crude oil prices.
- X_4: 4-month expiration crude oil prices.
- X_5: iShares U.S. Oil & Gas Exploration & Production ETF (IEO).[3]
- X_6: MVIS Global Oil Refiners Index (MVCRAK).[4]
- X_7: Seasonality extracted by numbering the months.

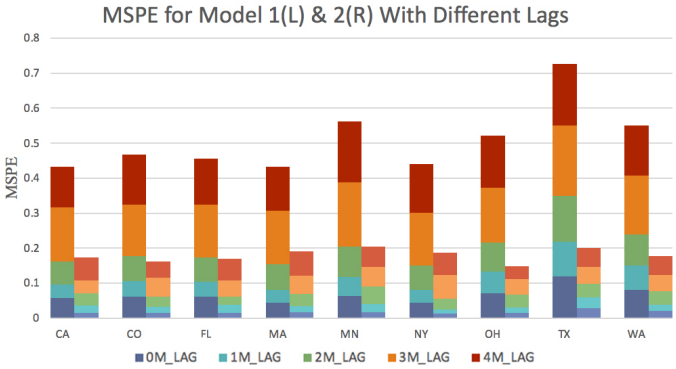

Fig. 4. Overall 63.8% reduction of MSPE from model 1; Texas' prediction has improved a lot. (Left Bar: Model 1, Right Bar: Model 2)

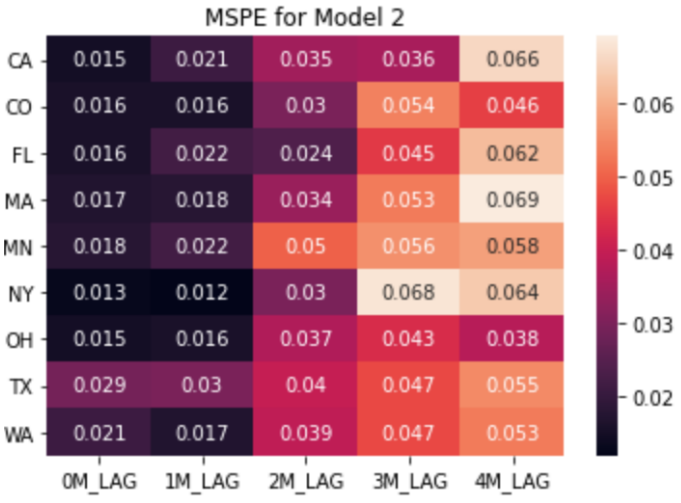

Fig. 5. Overall 63.8% reduction of MSPE from model 1; Texas' prediction has improved a lot.

[3] www.ishares.com
[4] www.mvis-indices.com

In addition, we add three additional variables to examine their impacts on the model. The first one is iShares U.S. Oil Gas Exploration Production ETF (IEO), in which the portfolio only contains spot equity products. The second one is the MVIS Global Oil Refiners Index (MVCRAK) which tracks the spot equity products of oil refinery companies. The third one is the month, indicating the gasoline price we are trying to predict is in which month of the year. By doing so, we aim to account for the potential seasonal patterns on gasoline prices. As the gasoline prices are closely related to the profitability of midstream and downstream oil companies, we believe these three additional variables would help us capture the changes reflected in the equity markets. From fig 4. above, which displays the MSPE for both model 1 and model 2, we can see that overall, the MSPE of the model has improved from model 1 by 63.8% on average, and 67.1% in small lag models, which means that the production and refinery index is useful in making gasoline price predictions. In addition, the model shows that it makes predictions well in both the concurrent and 1-month lag models. From the relative importance graph below (TABLE II), we can see that crude futures still have the highest level of predictability compared to the other factors. We calculate the percent of sample data that is explained by four maturities. With the increase in the lag period, crude oil futures' relative importance reduces, meaning that crude oil futures are less influential in long-term predictions.

TABLE II
MODEL 2 RELATIVE IMPORTANCE - TEXAS' PREDICTION ARE MORE CAPTURED BY MVCRAK (K); STATES WITH EXTREME WEATHER ARE MORE IMPACTED BY SEASONALITY (M); LONG-LAG MODELS' MOST IMPORTANT FACTOR TENDS TO BE THE CORRESPONDING EXPIRATION'S OIL FUTURES CONTRACT

Region	**Relative Importance**				
	0M_LAG	***1M_LAG***	***2M_LAG***	***3M_LAG***	***4M_LAG***
CA	13ikm42 [a]	13kim42	143kim2	31mik42	42mi1k3
CO	231kim4	213kim4	14kmi23	31ki4m2	4i1km23
FL	13kim24	1k4im32	14k3mi2	31imk42	42ikm13
MA	13ik4m2	23ikm14	143imk2	32ikm14	41imk23
MN	231imk4	241mik3	41mki32	34mik12	42mik13
NY	213im4k	231ik4m	1mik432	34mki12	42m1ik3
OH	213kmi4	23mki14	14mki23	1mi4k23	4m1ik32
TX	1kmi432	1kmi342	1k34mi2	2ikm413	4kmi123
WA	1kim342	1k3mi42	341kmi2	1m4ik32	42mik13

[*] 1,2,3,4: Time to expiry (in month) of oil futures.
[*] m: Month, k: MVCRAK, i: IEO.
[a] For model fitted with 0 month lag predicting CA gasoline prices, 1 month future prices have the most predictive power and 2 month future prices have the least predictive power.

Since IEO and MVCRAK are based on spot market products instead of futures, the model performs best in current weekly data rather than with a 1-month lag; this result is consistent with our expectation. Moreover, since Texas produces crude oil, the price fluctuations in crude oil do not play as much of a significant role in its gasoline price as other costs such as refinery factors. By inspecting the MSPE gain for IEO and MVCRAK, there is a 12.2% MSPE gain in the model; this matches the breakdown of 13% refinery cost as indicated by EIA[5]. In addition, Texas' MSPE is relatively at the same level as other states in this model compared to model 1. That coincides with our analysis in model 1 that additional parameters are needed to predict gasoline prices in Texas. Furthermore, the month also has some level of predictability, especially in states that have extreme weather, such as Minnesota and Ohio. It also coincides with the supply and demand analysis in economics. As in states with extreme weather, the supply of gasoline would be impacted by snowy roads and gasoline delivery delays to gas stations. The demand would also be impacted, as people tend to stay inside during extreme weather. Therefore, it is economically reasonable to conclude that the gasoline prices in such states would be more influenced by seasonality.

IV. GASOLINE PRICE MOVEMENT PREDICATION — PLANNING YOUR FUTURE GAS STATION VISIT WITH ADVANCED MACHINE LEARNING MODEL

A. Model Methodology, Feature Engineering, and Model Selection

1) Model Objectives:

Our models successfully predict whether the gasoline prices will go up or down in the next 1, 2, 3, and 4 weeks with accuracy as high as 70%. The models would enable consumers across the eight U.S. states to capture favorable pricing and plan their gas pump visits. Moreover, our models are able to provide insights into the magnitude of future changes in gasoline prices.

2) Feature Engineering:

We transform these objectives into two classification problems. Firstly, to predict the general direction of oil price changes, we assign negative percentage changes in gasoline price as class 0 and positive percentage changes in gasoline price as class 1. Secondly, to capture the magnitudes of gasoline price changes, we add more granular labels. Specifically, we assign all changes below the first quartile of all changes in the state as class 0, all changes between the first quartile and the third quartile as 1, and all changes above the third quartile as 2. We measure our models' successes based on the prediction accuracy of these labels. The labels are constructed respectively for predication periods 1, 2, 3, and 4 weeks into the future.

We construct two sets of independent variables from the crude oil future and crude oil spot prices data introduced in the previous parts of the paper.

The first sets of independent variables are the percentage differences between crude futures contracts that expired in 2, 3, 4 months and the futures contracts that expired in the current month. We believe this set of variables can capture the market's expectation for the future evolution of costs in the oil

[5]U.S. Energy Information Administration

industry and provide insights into how gasoline supplies will price the product at gas stations:

$$Crude\_Future\_D_1 = \frac{Future\_2M - Future\_1M}{Future\_1M}$$

$$Crude\_Future\_D_2 = \frac{Future\_3M - Future\_1M}{Future\_1M}$$

$$Crude\_Future\_D_3 = \frac{Future\_4M - Future\_1M}{Future\_1M}$$

The second sets of independent variables are the percentage changes between crude spot prices of the current week versus the spot prices 1, 2, 3, and 4 weeks ago. We believe this set of variables can effectively describe gasoline manufacturers' realized cost and explain their pricing behaviors.

$$Crude\_Spot\_D_1 = \frac{Spot_t - Spot_{t-1week}}{Spot_{t-1week}}$$

$$Crude\_Spot\_D_2 = \frac{Spot_t - Spot_{t-2week}}{Spot_{t-2week}}$$

$$Crude\_Spot\_D_3 = \frac{Spot_t - Spot_{t-3week}}{Spot_{t-3week}}$$

$$Crude\_Spot\_D_4 = \frac{Spot_t - Spot_{t-4week}}{Spot_{t-4week}}$$

3) Model Selection:

Firstly, we choose Logistic Regression as a simple benchmark model. As a classic classification model, Logistic Regression can capture a linear relationship between the independent variables. On the other hand, it cannot capture the non-linear relationships and long-term trends embedded in our time-series data.

Secondly, we apply Recurrent Neural Network (RNN) as an advanced machine learning tool that can take full advantage of the non-linear relationship in our data set. Moreover, the RNN model tracks the past evolution of our independent variables with proper units such as Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).

Eventually, we evaluate our models based on their overall accuracy in classifying future gasoline price movements.

B. Model 1 - Logistic Regression & Result Analysis in Texas

1) Model Configuration:

We set up the Logistic Regression model using feature engineering:

$$\pi_t(X) = \frac{\exp(\beta_0 + \sum_{i=1}^{3} \beta_i \cdot CF_{Di} + \sum_{i=1}^{4} \beta_{i+3} \cdot CS_{Di})}{1 + \exp(\beta_0 + \sum_{i=1}^{3} \beta_i \cdot CF_{Di} + \sum_{i=1}^{4} \beta_{i+3} \cdot CS_{Di})}$$

Where π is the probability that the oil price would increase in $t = 1, 2, 3, 4$ weeks.

We also set up a model for the 3-label classifiers and train it with the Sckit-Learn Logistic Regression function.

2) Result Analysis in Texas:

To illustrate our methodology and model performance, we use gasoline prices in Texas as an example. We train the model with data from an extended period, between the week of April 30, 1993, and the week of March 30, 2018. The validation period is between the week of April 6, 2018, and the week of December 25, 2020.

For the 3-label classifiers, we find the gasoline price changes in Texas are distributed in the following quartiles for our prediction horizons.

TABLE III
MAGNITUDE THRESHOLDS FOR DIFFERENT PREDICTION HORIZONS

	1 Week	***2 Weeks***	***3 Weeks***	***4 Weeks***
Major Decrease (0)	≤-0.9%	≤-1.7%	≤-2.5%	≤-3%
Mild Change (1)	Between	Between	Between	Between
Major Increase (2)	>0.9%	>1.7%	>2.5%	>3%

From the table below, we observe that the binary and 3-class classifiers produce strong prediction precision for future gasoline price changes in Texas. These findings help consumers to decide which week would be optimal to fill their tanks.

TABLE IV
BINARY AND 3-CLASS CLASSIFIERS HAVE HIGH VALIDATION PRECISION

	1 Week	***2 Weeks***	***3 Weeks***	***4 Weeks***
Binary Precision	72%	70%	72%	71%
3-Class Precision	69%	66%	67%	65%

C. Model 2 - RNN & Result Analysis in Texas

1) Model Configuration:

Instead of only looking at data available in the current period, RNN allows us to consider an extended period of crude oil data as a sequence. The data at each time step is fed into the hidden layer of the RNN model and carried forward to make predictions about the following time steps. Eventually, the final element of the sequence is fed into a Feedforward Neural Network, where a softmax function would eventually designate a class for the entire sequence.

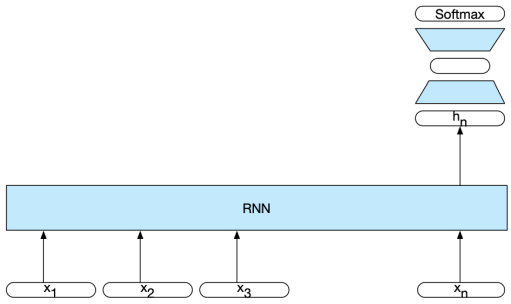

Fig. 6. RNN Architecture for Sequence Classification

Algorithm 2 Recurrent Neural Network

1: **Next Hidden Layer**

$$h_t = g(W_0 \cdot h_{t-1} + W_1 X_t)$$

2: **Classification**

$$y_t = softmax(W_2 \cdot h_t)$$

To better capture and remember long-term information, we use the Gated Recurrent Unit (GRU). The GRU unit includes a reset gate and an update gate that enables the unit to learn how to keep and improve its ability to handle long-term information. The RNN model includes 128 GRU units as a hidden layer to sufficiently capture the non-linear and long-term information. We use the Negative Log-likelihood function as the RNN's loss to train classification problems with multiple classes. To accelerate training, we use the Stochastic Gradient Descent Optimizer to minimize the loss function during training. Eventually, we add a 12-week lookback window such that our RNN will take data $X_{t-12}, X_{t-11}, ...X_t$ to train and perform predictions. That would allow us to take a quarter worth of data into account, which matches the business planning and reporting cycle of U.S. corporations.

2) Result Analysis in Texas:

The RNN models are trained on the same data set as the Logistic Regression model. Based on the results, we notice that the RNN models are better at predicting near-term gasoline price changes. We believe one reason for the performance difference is that the RNN model uses data from past periods. As a result, the RNN model predicts near-term, that are closer to the past periods, better than Logistic Regression. In general, the RNN models are also better at predicting the 3-class labels due to their ability to capture long-term and non-linear information.

TABLE V
BINARY AND 3-CLASS RNN HAVE HIGH VALIDATION PRECISION

	1 Week	*2 Weeks*	*3 Weeks*	*4 Weeks*
RNN Binary Precision	77%	72%	67%	65%
Logistic Binary Precision	72%	70%	72%	71%
RNN 3-Class Precision	75%	74%	70%	62%
Logistic 3-Class Precision	69%	66%	67%	65%

D. Implications for Consumers and Further Study

The models provide robust tools for consumers that could help them plan visits to gas pumps. According to the prediction of our classifier, consumers would be able to capture more favorable at-the-pump prices based on accessible data.

The Logistic Regression model stands out because it is interpretable and computationally efficient. On the other hand, the RNN models can capture more relationships. The RNN model will also allow future users to add more variables, such as alternative security indices, economic data, and geographical information.

The results for the seven other states are attached in Appendix A.

V. CRUDE OIL AND CO_2 CONCENTRATION

A. No Predictive Power

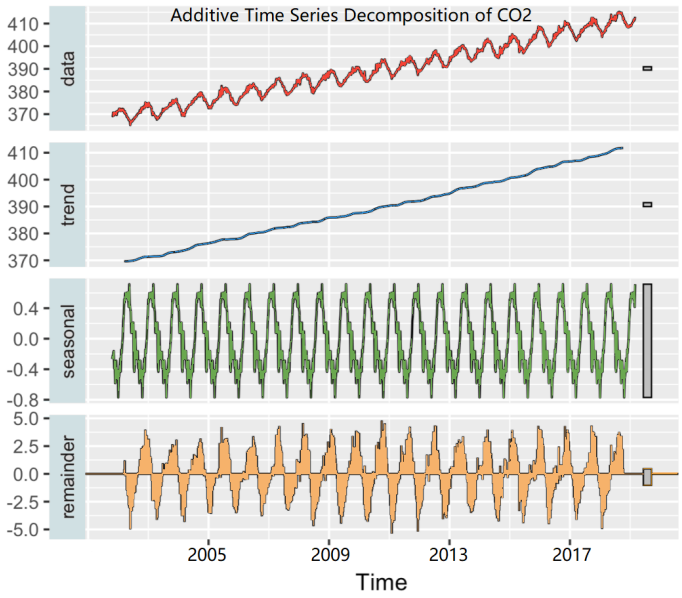

Fig. 7. CO_2 Concentration Decomposition with strong seasonality and trend

Decomposition of Additive Time Series:

$$Y_t = T_t + S_t + e_t \tag{4}$$

For the equation above, Y_t represents the CO_2 data, T_t represents the trend, S_t the seasonality, and e_t the remainder.

Figure 7 suggests an increasing trend and strong seasonality of the atmospheric CO_2 concentration since 2000. We find that the concentration level increases from September to June and then decreases. Moreover, based on the graph, we can see that the seasonality pattern and trend are stable over time and seem to have little relation to financial markets. We run another XGBoost model and predict with the same crude oil futures data and time lags. However, the MSPE is significantly large, as shown in Table VI, so we believe that crude oil futures prices do not have predictive power for the CO_2 concentration.

TABLE VI
MSPE (CO_2) USING ONLY FUTURES

	0M_LAG	*1M_LAG*	*2M_LAG*	*3M_LAG*	*4M_LAG*
CO_2	58.69	66.27	45.92	52.78	51.52

B. Improved Performance Using Additional Data

According to EPA[6], the main sources of greenhouse gas emissions by the economic sector in 2008 are Electricity (27%), Transportation (28%), and Industry (22%). Furthermore, the major component of global greenhouse gas emissions is CO_2 emissions from fossil fuels and industrial processes. It leads us to investigate the variables related to these areas and use additional data to build a model.

- Y: CO2 concentration level.
- X_1: U.S. total gasoline retail sales by refiner.
- X_2: Dow Jones Transportation Average Index (TRAN).
- X_3: Dow Jones U.S. Electricity Index (DJUSEU).
- X_4: Dow Jones Industrial Average (INDU).
- X_5: Seasonality extracted by numbering the months.
- X_6: Binary variable indicating trend (upward: Sept. - June; downward: June - Sept.).

Since the seasonal pattern recurs every one-year period, we believe that the effect for these economic sectors on CO_2 concentration is slow, and increase the time lags to 3 months, 6 months, 9 months, 12 months, and 18 months.

TABLE VII
MSPE (CO_2) WITH ADDITIONAL DATA

	3M_LAG	*6M_LAG*	*9M_LAG*	*12M_LAG*	*18M_LAG*
CO_2	2.87	2.26	2.42	2.54	4.75

As shown in Table VII, the MSPE improves significantly after we include additional variables in the model. Moreover, the MSPE with a lag of 6 months is relatively smaller than that with other time lags. We can now predict the CO_2 concentration level in the future period.

TABLE VIII
RELATIVE IMPORTANCE (CO_2) WITH ADDITIONAL DATA

	3M_LAG	*6M_LAG*	*9M_LAG*	*12M_LAG*	*18M_LAG*
CO_2	GTEImt	GETmIt	GEmIt	TGIEmt	GETImt

* G: Gasoline Sales, E: Electricity, I: Industrial, T: Transportation, m: Months, t: Trend

TABLE IX
RELATIVE IMPORTANCE OF GASOLINE SALES SHOWS STRONGEST PREDICTABILITY

	3M_LAG	*6M_LAG*	*9M_LAG*	*12M_LAG*	*18M_LAG*
Gasoline	68%	67%	66%	66%	65%

Table VIII shows the relative importance of all variables used in the model. It suggests that the variable "G" which represents the gasoline sales is the most important factor. From table IX, we can see that gasoline sales explain over 50% of CO_2 concentration. Since the sales data indirectly reflects the consumption level of gasoline in the U.S., the result indicates that the increasing CO_2 concentration level is highly related to the gasoline consumption, which coincides with our expectation.

[6]U.S. Environmental Protection Agency

VI. CONCLUSION

In conclusion, crude oil futures have predictive power for future gasoline prices as it is the largest influencing factor in forecasting. Although gasoline is always treated as an inelastic product, it does show price impacts on changing demand and supply according to seasonal and cost measurement.

We find that past crude oil future and spot price data have significant predictive power over the future movement of gasoline prices from our classification analysis. By applying machine learning techniques such as Logistic Regression and Recurrent Neural Network, we create tools that can assist everyday consumers in shopping for better prices at gas stations across the U.S.

With weak predictive power of crude oil futures prices to CO_2 concentration, our new model captures gasoline sales in the U.S. as the most dominant predictor, with over 60% explanatory power to CO_2 concentration. Thus, reducing gasoline sales and consumption should be the number one priority in making carbon policies.

VII. FUTURE WORK

In addition to the parameters that we add to explore the gasoline markets, there are still other long-term economic factors that may be impacting the gasoline price, such as tax rates in different states, as well as geopolitical events. Adding these influences into a gasoline price prediction model needs supercomputing and more complicated deep learning techniques to extract useful information.

The RNN models we build are Black-Box models that have poor interpretability. Although the classifier is able to generate useful predictions, it is hard to understand the economic intuitions behind the results, which makes it challenging to improve the model and minimize errors. In the future, we would like to perform more economic analysis on the variables and find additional data that better describe the differences between states. For instance, the oil future data we used is about contracts delivered to Oklahoma. That could partially explain why the models perform less optimal in Massachusetts, Minnesota, and New York, whose oil supplies could come from Canada or North Sea imports.

REFERENCES

[1] Jurafsky D. Martin H. M., "Speech and Language Processing, Chapter 9", 3rd ed.,2020, pp. 2.

[2] Friedman J., Hastie T. Tibshirani R., "The Elements of Statistical Learning:Data Mining, Inference and Prediction", 2nd ed.,2008, pp. 640.

[3] Chen T. Guestrin C., "XGBoost: A Scalable Tree Boosting System", 2016.

[4] Roll R., "Orange Juice and Weather", *The Ameri.can Economic Review*, Vol. 74, No. 5,Dec. 1984, pp. 861-880.

[5] EPA. Sources of greenhouse gas emissions, 2020, December 04, Retrieved February 1, 2021, from https:// www.epa.gov/ ghgemissions/ sources-greenhouse-gas-emissions.

[6] EIA. "Gasoline explained:Factors affecting gasoline prices", February 06, 2020, Retrieved February 1, 2021, from https://www.eia.gov/ energyexplained/ gasoline/ factors-affecting-gasoline-prices.php.

[7] IPCC, "Climate Change 2007: Mitigation of Climate Change", 2007, pp.169-250.

APPENDIX

A. Result Comparison Between Eight States in the U.S.

	1 Week	***2 Weeks***	***3 Weeks***	***4 Weeks***
TX RNN Binary	77%	72%	67%	65%
TX Logistic Binary	72%	70%	72%	71%
TX RNN 3-Class	75%	74%	70%	62%
TX Logistic 3-Class	69%	66%	67%	65%
MA RNN Binary	68%	72%	59%	55%
MA Logistic Binary	73%	74%	65%	71%
MA RNN 3-Class	46%	51%	68%	53%
MA Logistic 3-Class	44%	46%	54%	57%
CA RNN Binary	76%	69%	65%	67%
CA Logistic Binary	74%	73%	66%	72%
CA RNN 3-Class	58%	56%	55%	62%
CA Logistic 3-Class	55%	49%	53%	62%
CO RNN Binary	74%	69%	66%	67%
CO Logistic Binary	72%	72%	72%	71%
CO RNN 3-Class	56%	56%	56%	61%
CO Logistic 3-Class	61%	58%	61%	61%
FL RNN Binary	73%	64%	62%	62%
FL Logistic Binary	74%	68%	66%	64%
FL RNN 3-Class	68%	57%	57%	60%
FL Logistic 3-Class	62%	52%	54%	60%
MN RNN Binary	66%	63%	49%	46%
MN Logistic Binary	68%	66%	63%	61%
MN RNN 3-Class	56%	54%	54%	54%
MN Logistic 3-Class	52%	58%	59%	56%
OH RNN Binary	72%	58%	66%	63%
OH Logistic Binary	66%	63%	63%	64%
OH RNN 3-Class	75%	70%	65%	65%
OH Logistic 3-Class	73%	69%	61%	64%
NY RNN Binary	70%	69%	65%	58%
NY Logistic Binary	68%	70%	68%	60%
NY RNN 3-Class	54%	48%	52%	56%
NY Logistic 3-Class	46%	47%	50%	53%