Communications of the IIMA

Manuscript 1440

Forecasting Gasoline Price with Time Series Models

Xin James He

Forecasting Gasoline Price with Time Series Models

Xin James He
Fairfield University
1073 North Benson Road, Fairfield, CT 06824, USA

ABSTRACT

This research forecasts the gasoline price in U.S. and analyzes its managerial implications by means of both univariate and multivariate time series forecasting models. Gasoline price forecast is among the most difficulty time series variables during its importance to the economy and extremely volatile nature. The average regular gasoline price in U.S. reached $5.06 per gallon in the month of June 2022, as opposed to $3.41 at the beginning of 2022, a 48% increase. While gasoline prices had been rising over the first half of 2022 due to supply chain disruptions as a result of global Covid 19 lockdowns and the Russia invasion of Ukraine since February 24, 2022, they then surprisingly came down in the second half of 2022 to $3.85 in November 2022, which has caught many consumers and business organizations off-guard. Both univariate time series forecasting models, such as exponential smoothing and autoregressive integrated moving average, and multivariate time series forecasting models, such as time series regression models, are used in this research with the data for the period January 2002 through November 2022. We find that the time series regression model with trend, season, GDP, CPI, and crude oil price turns out the be the best forecasting model both on the training data and the testing data, even with the testing data containing significant turning points. Managerial implications and future research directions are also discussed.

Keywords: Time Series Forecast, Gasoline Price, R Studio, ARIMA Models, Time Series Regression, Model Accuracy and Validation

INTRODUCTION

This research tries to forecast gasoline prices and analyze their managerial implications by investigating both univariate and multivariate time series forecasting models. Gasoline price forecast is among the most difficulty time series variables during its importance to the economy and extremely volatile nature. The average regular gasoline price in US had reached $5.06 per gallon in the month of June 2022, as opposed to $3.41 at the beginning of 2022, a 48% increase. While gasoline price had been rising over the first half of 2022 due to supply chain disruptions as a result of global Covid 19 lockdowns and the Russia invasion of Ukraine since February 24, 2022, it equally surprisingly came down in the second half of 2022 to $3.85 in November 2022, which has caught many consumers and business organizations off-guard scrambling to deal with the extremely volatile energy prices. Higher gasoline prices would not only lead to higher consumer prices, but also have a negative effect on consumer purchasing power and may eventually slow down the economy, or even trigger a recession. Since gasoline prices are random variables that are measured over time, univariate time series forecasting models will be considered as baseline models, such as exponential smoothing with trend and seasonality (ETS), linear regression with trend and seasonality, and autoregressive integrated moving average (ARIMA). Moreover, multivariate time series forecasting models will also be used to deal with extremely volatile gasoline price and to improve the forecasting accuracy, with such predictor variables as GDP, CPI, crude oil price, unemployment rate, and federal funds rate, with the gasoline price as the target variable. Managerial implications and future research directions are also discussed.

LITERATURE REVIEW

Time series forecasting models for crude oil and gasoline prices can be divided into two competing categories: univariate time series forecasting models and multivariate time series forecasting models. Univariate time series models use only a panel of historical data to produce forecasts, under the assumption that the historical pattern continues in the future. However, when this assumption does not hold true, especially the forecasting period demonstrates significant turning points, the forecasting accuracy will be severely compromised (Lusk, 2019).

He (2018) reports that simple moving average and simple exponential smoothing models such as MA (2) and SES ($\alpha = 0.9$) can provide reasonably acceptable forecasting accuracy with minimum computational complexity.

In general, simple moving average (MA) and simple exponential smoothing (SES) are the most commonly used forecasting methods for time series data in U.S. government statistics and stock prices (Huntington, 1994; Abramson and Finizza, 1995). Hyndman and Athanasopoulos (2021) introduced exponential smoothing with trend and seasonality (ETS), along with Holt-Winters' and damped exponential smoothing forecasting models. Laung-Iem and Thanarak (2021) try to use time series decomposition method based on the data from 2007 – 2016 to make a 25-year long term forecast on diesel prices for 2017 - 2036. Autoregressive integrated moving average (ARIMA) models are among the most prominent univariate time series models, where its autocorrelation function (ACF) and partial autocorrelation function (PACF) are used to help select data driven model parameters (Ord et al, 2017). When it is done correctly, ARIMA models can provide very accurate forecasting results, especially for short-term time series data (Xiong et al, 2013; Box et al, 2015; Dritsaki et al, 2021).

Univariate machine learning forecasting models, such as support vector regression (SVR) and artificial neural network (ANN), have been widely used to forecast crude oil and gasoline prices (Basak, Pal, and Patrianabis, 2007). Xie et al (2006) find that SVR outperforms ARIMA based on monthly spot prices of West Texas Intermediate (WTI) crude oil from January 1970 to December 2004. Sehgal and Pandey (2015) concede after reviewing various artificial intelligence methods, including SVR and ANN, that the existing literature is very far from any consensus about a reliable forecasting model regarding crude oil prices. According to He (2018), while it may marginally improve the forecasting accuracy over ARIMA models, the SVR models are not only computationally more complex among all the forecasting models analyzed, but also have the potential of model overfitting due to the fact that there are too many parameters to train the model: cost, gamma, and epsilon. In addition, an SVR model cannot be used to test the model accuracy on the testing data the same way as in an ARIMA model since it does not provide a list of model parameters, which also makes the economic or business interpretation very difficult. Yao and Wang (2021) use a combined Long Short-Term Memory (LSTM) network and grey prediction model (GM) to forecast WTI crude oil prices for the period Jan. 1986 through Jan. 2020. However, the main focus of this research is to decompose the time series into multiple sequences and then suggests various prediction models to improve forecasting accuracy.

Multivariate Time Series Models – Multiple linear regression models (LM) with multivariate time series can include trend and seasonality in addition to predictor variables (Hyndman and Athanasopoulos, 2021). Since regression analysis is able to explore the interconnections between gasoline prices and other independent variables such as GDP and CPI, it was used to help interpret models, especially for managerial and policy implications (Chinn, LeBlanc, and Coibion, 2005). Breiman

et al (1984) proposed Classification and Regression Tree (CART) in their book *Classification and Regression Tree*. They defined the regression model as Tree Structured Regression to differentiate it from other regression methods, where the training set is partitioned by a sequence of binary splits into terminal nodes. In each terminal node, a numerical value will be generated as the predicted value at each leaf node. Chen and He (2019) test forecasting model accuracy on 1992-2017 West Texas Intermediate (WTI) oil price data by comparing Classification and Regression Tree (CART) - Random Forest models with multiple linear regression and ARIMA models. However, a major drawback of a machine learning model is that it is difficult to interpret the process in meaningful statistical or business perspectives due to the fact that it relies on a high dimensional space via a nonlinear function, which makes managerial and policy implications more difficult.
According to U.S. Energy Information Administration (http://www.eia.gov) dated November 2022, the retail price of gasoline includes four main components: the cost of crude oil (55%), federal and state taxes (14%), distribution and marketing costs and profits (18%), and refining costs and profits (13%). Since taxes and profits are usually related to overall econoimic conditions, it is logical to include such predictor variable such as crude oil price, along with consumer price index (CPI) and gross domestic product (GDP), to forecast the gasoline price.

METHODOLOGY

Forecasting Models

The following univariate time series models are scrutinized for the gasoline price forecasting models: a) simple exponential smoothing (SES), b) the Holt-Winters' seasonal models (Holt), c) the Holt-Winters' damped models (Damped), d) exponential smoothing with trend and seasonality (ETS), and e) autoregressive integrated moving average (ARIMA) models, The Holt method comprises the forecast equation and three smoothing equations, the level ℓt, the trend bt, and the seasonal component st, with corresponding smoothing parameters α, β, and γ. The Damped method is often in the form of a damped additive trend and multiplicative seasonality in order to improve forecasting accuracy. The ETS model is a generalized form of exponential smoothing with trend and seasonality, where E stands for error, T for trend, and S for seasonality. The ARIMA model is often used as a benchmark in terms of forecasting accuracy against machine learning models.

Forecasting accuracy and model validation will be assessed by comparing such metrics as RSquared, standard error (St-Error), square root of mean squared error (RMSE), and mean absolute percentage error (MAPE).

Data Collection

In this research, we collect the data for the period January 2002 through November 2022 both for univariate time series and multivariate time series forecasting models. The monthly U.S. gasoline prices (GasPrice) is used as the time series variable for univariate forecasting models, and as the target variable for multivariate time series forecasting models, along with five predictor variables. One of the predictor variables, West Texas Intermediate (WTI) spot prices (OilPrice), is collected from the U.S. Energy Information Administration (http://www.eia.gov). The other four predictor variables - U.S. Gross Domestic Production (GDP), U.S. Consumer Price Index (CPI), U.S. Unemployment Rate (Unemploy), and U.S. Federal Reserve Interest Rate (FedRate) - are collected from the U.S. Federal Reserve (https://fred.stlouisfed.org).

Since there has not been a consensual time series forecasting model accuracy and validation strategy among out-of-sample validation, prequential, or k-fold cross-validation approaches (Cerqueira et al, 2020), we focus out attention on the out-of-sample method in this research for model accuracy and validation.
Figure 1 shows the monthly U.S. regular gasoline prices from January 2002 through November 2022, with a significant volitivity since early 2020, especially since the beginning of 2022 following the Russia invasion of Ukraine. In addition, the gasoline price peaked around June 2022, forming an inverse V shape for the year 2022. In this research, we designate January 2002 – December 2021 as the training period and the January - November 2022 as the out-of-sample testing period. It is seen from Figure 1 that the testing period demonstrates extremely volatile turning points, which is often considered as one of the worst-case scenarios for time series forecasting accuracy and model validation (Lusk, 2019).

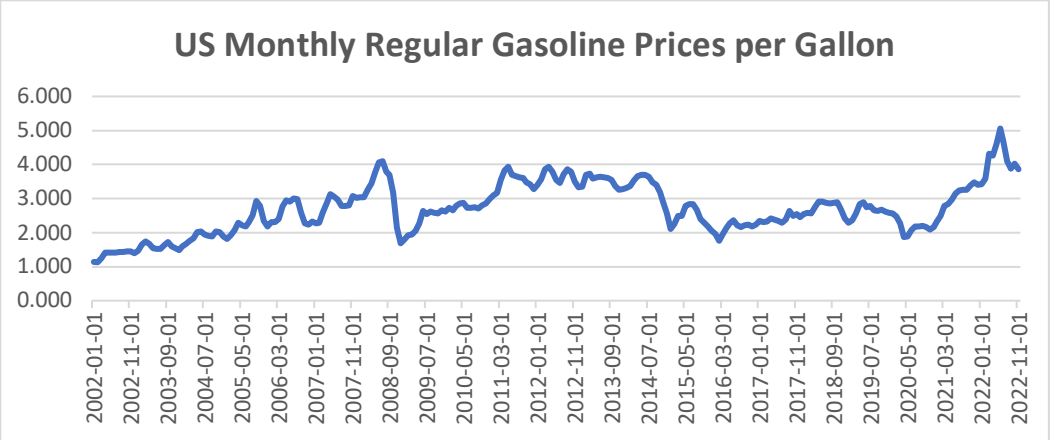

Figure 1. U.S. Monthly Regular Gasoline Prices Per Gallon 2002 –2022

Figure 2 illustrates the monthly WTI crude oil prices (OilPrice) from Jan. 2002 through Nov. 2022, indicating its major impact on gasoline price (GasPrice) but not necessarily in the same directions all the time. The OilPrice, GDP, CPI, Unemploy, and FedRate will be analyzed as predictor variables for multivariate forecasting models.

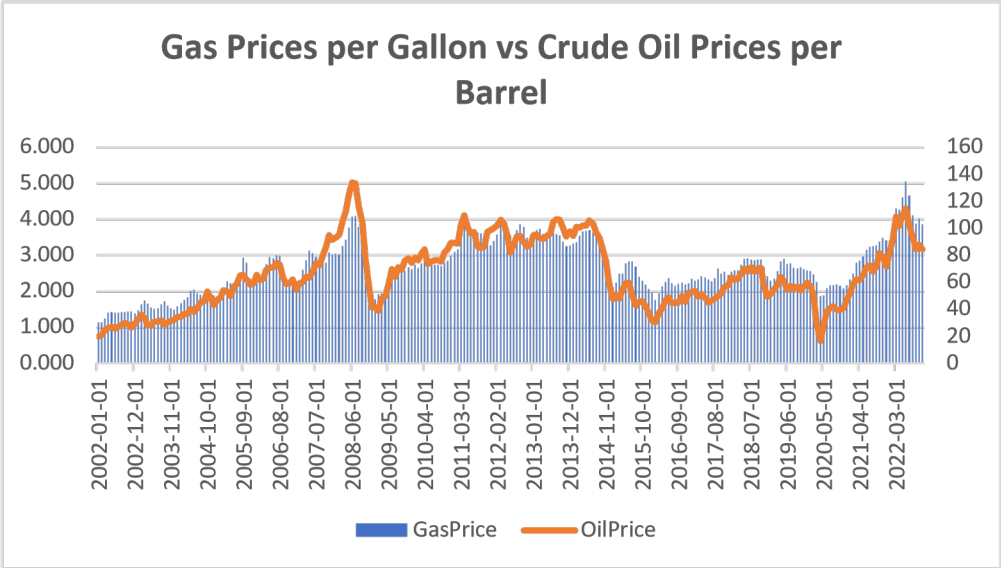

Figure 2. Regular Gas Prices vs WTI Crude Oil Prices 2002 –2022

FORECASTING MODEL DEVELOPMENT

R-Studio is deployed as the main platform for all forecasting model development, including seasonality analysis, univariate models, and multivariate time series models.

Seasonality Analysis

We analyze the seasonality of the gasoline price to determine if it is necessary to take into consideration of the seasonal components and to reveal the seasonal factors if any. Figure 3 depicts the seasonally adjusted gasoline prices in blue, the trend in red, and the actual price in gray. Table 1 provides the additive seasonal factors, where the month of May has the most positive factor of $0.1167 among the seven peak summer months (March – September) and the month of December has the most negative factor of -$0.1706 among the five winter valley months (November – February).

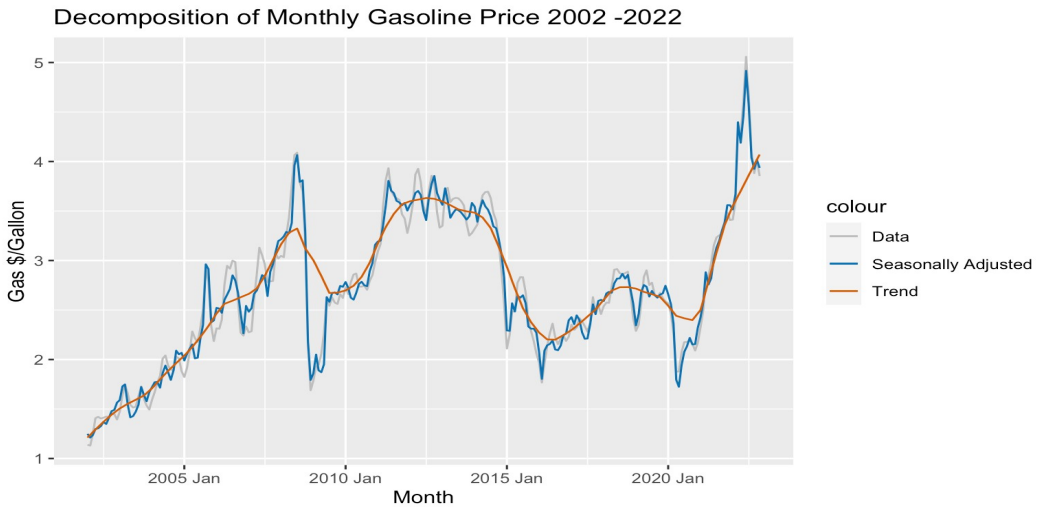

Figure 3. Gasoline Price Trend and Seasonality Analysis

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
S Factor	-0.1101	-0.0807	0.0059	0.1090	0.1167	0.0769	0.0433	0.0747	0.0099	-0.0296	-0.0435	-0.1706

Table 1. Gasoline Price Seasonal Factors

Univariate Forecasting Models

We first fit the train data on five most commonly used univariate time series forecasting models using R Studio: simple exponential smoothing (SES), Holt-Winters' (Holt), Damped Holt, ETS, and ARIMA. Table 2 shows that the ARIMA model performs the best on the train data with the smallest RMSE of 0.142, whereas the Holt model performs the best on the test data with a RMSE of 0.697. Please note that it seems counter intuitive that the ARIMA falls from the best train model to the worst on test data, which is likely due to the fact that the pattern on the train data does not continue on the test data with significant turning points. Moreover, the RMSRs on the test data are noticeably larger that these on the train data, which leads us to inspect the model performance graphically.

Model	Parameters	Train Data			Test Data		
		RMSE	MAPE	Rank	RMSE	MAPE	Rank
ARIMA	(2,1,2)	0.142	4.18	1	1.103	22.6	5
ETS	(M,N,A)	0.156	4.35	2	0.746	13.9	2
Damped	a, b, f	0.162	4.96	3	0.945	18.6	4
SES	a	0.169	4.74	4	0.884	17.1	3
Holt	a, b	0.172	4.95	5	0.697	11.7	1

Table 2. Univariate Forecasting Model Accuracy and Validation

Figure 4 provides the graphic comparison of the best model on the train data from Tables 2, ARIMA, and the best of model on the test data, Holt. It is seen from Figure 4 that both univariable forecasting models are almost equally accurately to fit the pattern of the train data, but they both miss the pattern of the test data in a big way: the ARIMA model (the red line) incorrectly pointing to a downward trend for the entire holdout period (Jan. -Nov. 2022), whereas the Holt model (the blue line) correctly predicting the upward trend but failing to catch the initial steep upward trend (Jan. – Jun. 2022) and then the dramatic downward trend (Jul. – Nov. 2022). It is no surprise to the time series discipline that even a benchmark univariate forecasting model such as ARIMA may often underperform when the out-of-sample data contains significant turning points, which lead us to the development of multivariate forecasting models.

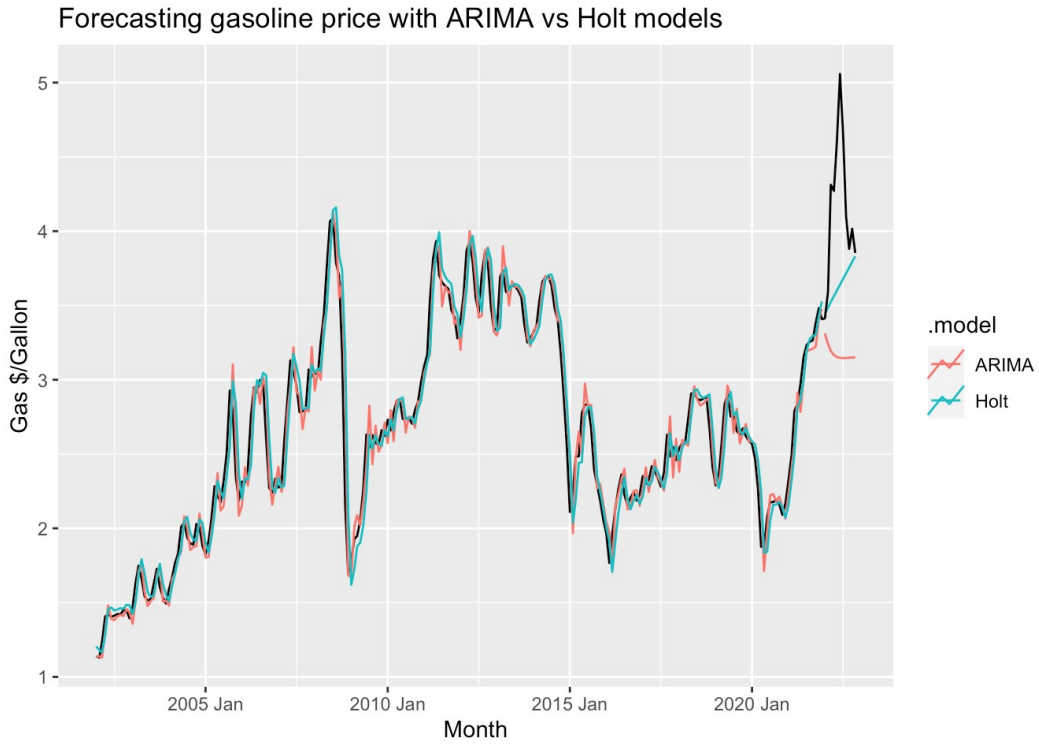

Figure 4. Univariate Forecasting Models Comparison: ARIMA vs Holt

Multivariate Forecasting Models

Figure 5 is the correlation matrix among all six variables with five interesting observations, where (***) indicating a p-value < 0.001. First, there is a strong positive correlation (0.927) between GasPrice and OilPrice. Second, there is also a strong positive correlation (0.547) between GasPrice and CPI. Third, CPI has a strong positive correlation (0.318) between CPI and OilPrice, but a strong negative correlation (-0.327) between CPI and FedRate. Figure 3 that both GDP and CPI are somewhat related to the gasoline prices, but the relationship is not as strong as that between the gasoline prices and the crude oil prices.

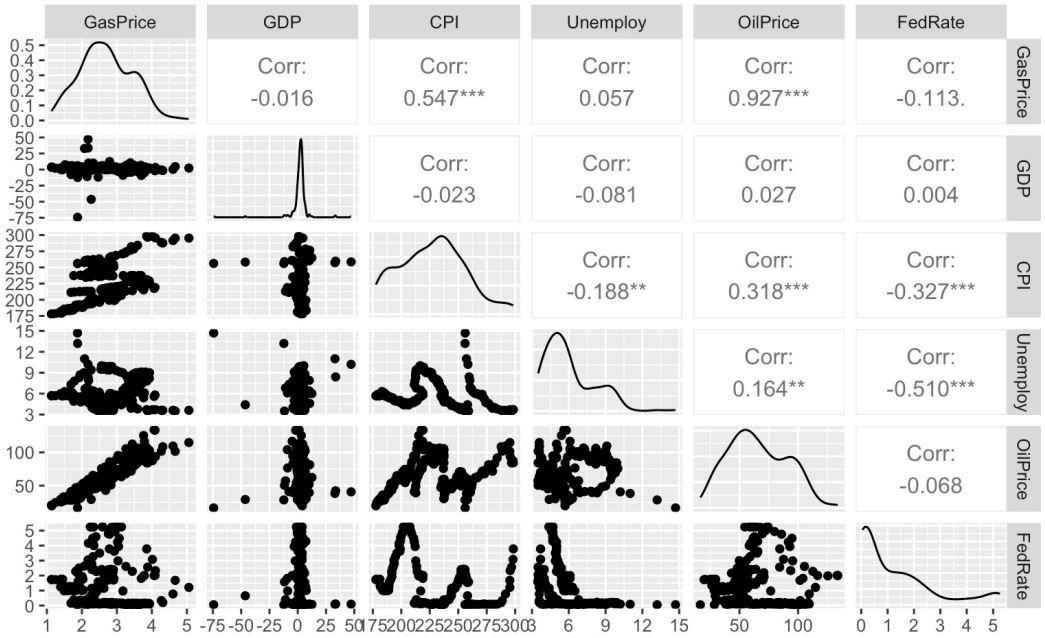

Figure 5. Correlation Matrix among Six Variables

In light of seasonality as shown in Figure 3 and Table 1, and the strong correlation (0.927) between GasPrice and OilPrice and as well as the significant correlation (0.547) between GasPrice and CPI as shown in Figure 2 and in Figure 5, we now focus our attention on multivariate time series regression models.

Table 4 depicts the multivariate time series regression models with GasPrice being the target variable and Y-intercept statistically significant at 0.001 (***) for all five models. LM0, for example, is the full regression model with all five predictor variables, along with time variable t for Trend and 11 monthly dummy variables for Season. However, since two of predictor variables, Unemploy and FedRate, are not statistically significant, which will be eliminated from further consideration. Consequently, the remaining four regression models are: LM1 – after eliminating both Unemploy and FedRate from LM0, LM2 - with only Trend, Season, and OilPrice, LM3 – with only Trend and Season, and LM4 with only OilPrice. Since LM0 has two predictor variables, Unemploy and FedRate, statistically insignificant and LM4 has a very low R-Squared value of 0,155, we eliminate them from further consideration.

Model	Trend	Season	GDP	CPI	Unemploy	OilPrice	FedRate	F
LM0	Y***	Y*	Y*	Y**	Y	Y***	Y	236***
LM1	Y***	Y*	Y*	Y**		Y***		268***
LM2	Y***	Y*				Y***		264***
LM3						Y***		1666***
LM4	Y***	Y						3.47***

*** = significant level 0.001 ** = significant level 0.01 * = significant level 0.05

Table 4. Regression Models Parameter Details

Table 5 shows the forecasting accuracy and validation results with both training and testing datasets in terms of R-Squared, St-Error, RMSE, and MAPE for regression models LM1, LM2, LM3, and LM4. It is seen from Table 4 that LM1 is the best model with the lowest RMSE and MAPE both for Training and Testing datasets, which ranks the first among all regression models. It is also worth noting that LM3, the model with OilPrice as the sole predictor variable, falls to the third place

Model	R-Squared	St-Error	Training		Testing		Rank
			RMSE	MAPE	RMSE	MAPE	
LM1	0.947	0.162	0.156	4.53	0.208	4.84	1
LM2	0.938	0.174	0.169	4.97	0.484	10.27	2
LM3	0.874	0.242	0.241	8.219	0.739	16.29	3

Table 5. Regression Models Accuracy and Validation Comparison

LM1: Y = b0+T+S2+S3+S4+S5+S6+S7+S8+S9+S10+S11+S12+GDP+CPI+OilPrice

Y=-4.551-008T+.026S2+.12S3+.183S4+.254S5+.237S6+.193S7+.194S8+.194S9
+.128S10+.06S11+.004S12-.003GDP+.03CPI+.022OilPrice (1)

LM3: Y = b0 +Oil Price = .952 + .026OilPrice (2)

It is worth noting that crude oil price alone is not sufficient to predict gasoline price as in LM3 of Eq. (2), as opposed to the much more accurate forecasting model in LM1 of Eq. (1) with trend, seasonality, GDP, CPI, and crude oil price as predictor variables.

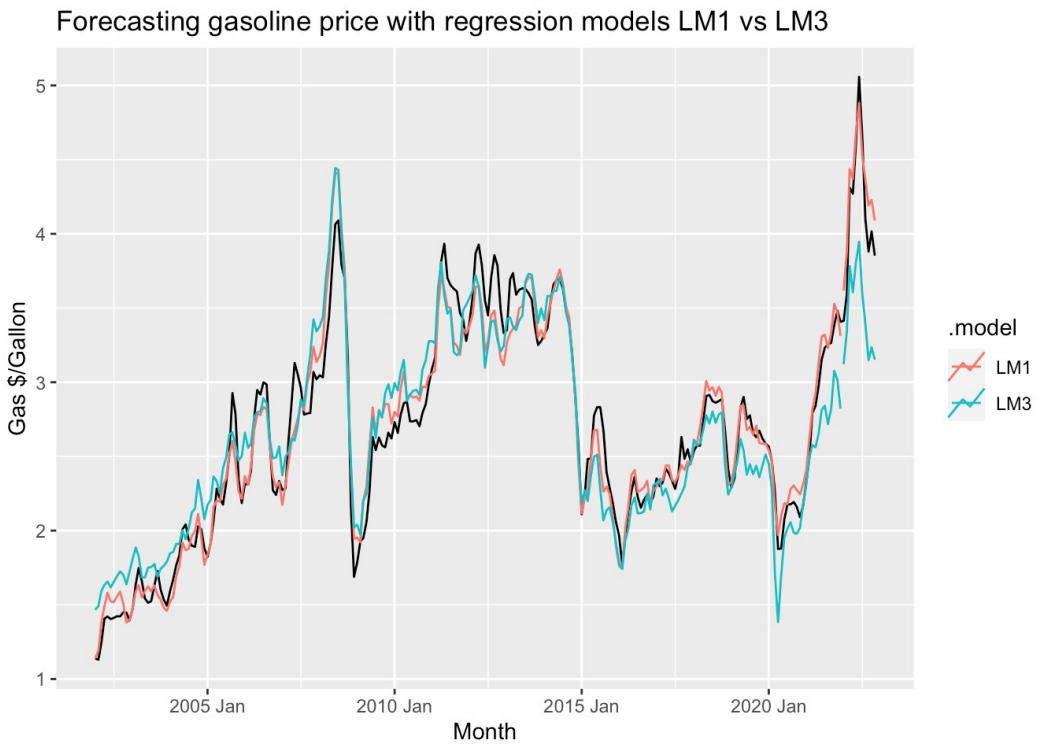

Figure 6. Regression Models Performance Comparison: LM1 vs LM3

Figure 7 provides a graphic view of the behavior of the final forecasting model comparison. It is seen from Figure 7 that the forecast value of the time series regression model, LM1, on the test data almost overlaps the actual observation in blue, as opposed to that of the ARIMA model in red, and the Holt model in green. Specifically, the red line from the ARIMA model seems to go in the opposite direction from the actual data, whereas the green line from the Holt model seems to be in the right direction but trails way behind that from LM1.

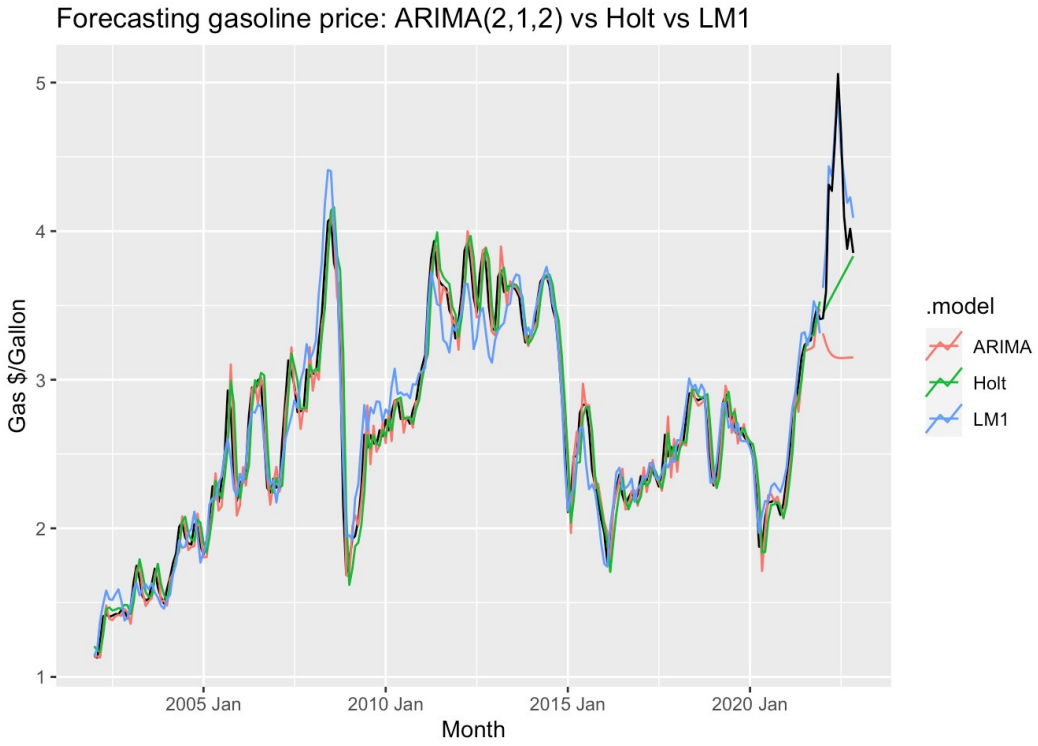

Figure 7. Forecasting Model Comparison: ARIMA vs Holt vs LM1

CONCLUSION

The ramification that the LM1 model outperforms all the models we investigated is significant. First, it confirms that the crude oil price plays a major role (55%) (http://www.eia.gov) in determining the gasoline price, but the crude oil price alone is not sufficient to predict the gasoline price as shown in LM3 model and Eq. (2), where CPI and GDP also affecting the gasoline prices. Second, the ARIMA model along with others seems to perform well on the train data, but turns out to be inferior when the test data contains significant turning points that do not follow the same pattern as the train data. Third, it will be interesting to investigate if the LM1 model can be a reliable forecasting model for gasoline price in the future or even be generalized to other energy related time series forecasting models.

REFERENCES

Abramson, B., Finizza, A. (1995). Probabilistic Forecasts From Probabilistic Models: A Case Study. *International Journal of Forecasting*, 11 (1), 63-72

Baghestani, Hamid, and Bley, Jorg (2020). Do Directional Predictions of US Gasoline Prices Reveal Asymmetries? *Journal of Economics and Finance*, 44, 348-360

Basak, Debasish, Pal, Srimanta, and Patrianabis, Dipak C. (2007). Support Vector Regression. *Neural Information Processing – Letters and Reviews*, 11 (1), 203-224

Box GEP, Jenkins GM, Reinsel GC, Ljung GM (2015). Time Series Analysis: Forecasting and Control. London: Wiley

Breiman, Friedman, Olshen and Stone (1984). Classification and Regression Trees. Belmont, Calif.: Wadsworth International Group

Cerqueira, V., Torgo, L. & Mozetic, I. (2020). Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods. *Machine Learning*, 109 (11), 19972028

Chen, Engu and He, Xin James (2019). Crude Oil Price Prediction with Decision Tree Based Regression Approach. *Journal of International Technology and Information Management*, 27 (4), 1-16

Chinn, MD, LeBlanc, M., and Coibion, O. (2005). The Predictive Content of Energy Futures: An Update on Petroleum, Natural Gas, Heating Oil and Gasoline. Technical Report, National Bureau of Economic Research