Fuelonomics: Analyzing Gas Price Dynamics through Machine Learning

Yimo Shen

2023-06-04

## Introduction

The fluctuating gas prices in the United States have a significant impact on the economy, transportation sector, and the daily lives of individuals. Understanding the factors that influence gas prices and accurately predicting their future values can provide valuable insights for policymakers, businesses, and consumers alike. In this project, we aim to investigate the determinants of gas prices in the U.S. and develop a predictive model using machine learning methodologies.

## Data

The project follows a structured approach, beginning with data collection. The project leverages data from two main sources: Statistical Review of World Energy published by BP, a reputable source for global energy economics, and the U.S. Energy information Administration. The first source provides data on energy prices, consumption, and production for a variety of sources of energy, like oil, natural gas, renewables etc. The second source provides data on prices of gasoline specifically. We analyzed data from the time period 1965-2021. Data from these sources provide a comprehensive information on global energy markets.

The initial dataset provided by BP contained data on various energy metrics from different countries. To narrow down the focus specifically to the U.S., all the relevant U.S. metrics were combined into a single dataset. One important step in the data preparation process was addressing missing values. To fill in these missing values, alternative sources such as government websites were consulted. By cross-referencing and validating the missing values from different sources, the dataset’s completeness was improved, ensuring that crucial information was not overlooked or excluded from the analysis.

Furthermore, efforts were made to enhance the quality of the dataset by eliminating columns that exhibited perfect correlation or collinearity. This involved removing columns that represented the same metric but with different units and identifying and eliminating subcategories that added up to another column. This step reduced redundancy and improved the model’s robustness by avoiding multicollinearity issues, ensuring that each variable contributed unique and meaningful information to the analysis.

Overall, the data preparation process involved consolidating the relevant U.S. metrics, filling in missing values through extensive research, and removing redundant columns. These steps were crucial in creating a refined and reliable dataset for further analysis, setting the foundation for investigating the determinants of U.S. gas prices accurately and drawing meaningful insights from the data.

## ML

The next step of our project is to use Machine Learning algorithms to analyze gas prices data. Machine learning offers a powerful approach to investigate the determinants of gas prices in the United States and develop accurate predictive models. By leveraging vast amounts of historical data, machine learning algorithms can identify key features that influence gas prices, ultimately providing valuable insights for decision-making. This project’s analysis consists of two parts: feature selection and prediction. Two popular algorithms for feature selection in this context are a simple linear regression and Lasso regression. RFE iteratively removes less relevant features, while Lasso regression performs variable selection by introducing a penalty term. Once the relevant features are identified, two effective algorithms for gas price prediction are Random Forest Regression (RFR) and Neural Network. RFR creates an ensemble of decision trees to make accurate predictions, while Neural Network models use interconnected layers of artificial neurons to learn complex patterns and relationships in the data. By comparing these algorithms and selecting the one with the best performance, machine learning can enable the investigation of gas price determinants and develop robust predictive models for future gas price predictions in the U.S.

### Feature Selection

Feature selection can bring several benefits, such as reducing overfitting, improving model training and inference speed, enhancing interpretability, and simplifying the underlying model.

#### Simple Linear Regression

The first feature selection algorithm our group decided to use was a simple linear regression. Linear regression can be a valuable tool for feature selection and exploring the determinants of U.S. gas prices. By fitting a linear model to the data, linear regression allows us to identify the relationships between the target variable (gas prices) and various predictor variables. Through the process of model building, linear regression estimates the coefficients of the predictor variables, indicating their impact on gas prices. This estimation process inherently performs feature selection by assigning higher coefficients to variables that have a stronger influence on gas prices, while reducing the importance of variables with little impact. By analyzing the coefficients, we can identify the key determinants of gas prices and understand their respective magnitudes and directions of effect. Linear regression also provides valuable insights into the statistical significance of the predictor variables, helping to distinguish between significant factors and random fluctuations. Thus, linear regression serves as a versatile tool for feature selection and uncovering the determinants of U.S. gas prices, enabling researchers to gain a deeper understanding of the factors driving price fluctuations in the market.

Based on the results of the linear regression model on gas prices in the U.S., we can interpret the coefficients as follows:

* **“log(data$oil\_production\_bbl)”**: This variable has a negative coefficient (-0.64533) with a highly significant p-value (p < 0.001). It suggests that an increase in oil production is associated with a decrease in gas prices, assuming all other variables are held constant.
* **“log(data$oil\_crude\_oil\_prices)”**: This variable has a positive coefficient (0.30284) and a highly significant p-value (p < 0.001). It indicates that higher crude oil prices are associated with higher gas prices.
* **“log(data$oil\_consumption\_bbl)”**: The coefficient for this variable is positive (0.30494), but it has a relatively high p-value (0.595555). This p-value suggests that the relationship between oil consumption and gas prices may not be statistically significant in this model.
* **“log(data$primary\_energy\_consumption\_per\_capital)”**: This variable has a negative coefficient (-1.99545) with a p-value of 0.046067 (\*). The p-value indicates that this variable is marginally significant in explaining gas prices. The negative coefficient suggests that higher primary energy consumption per capita is associated with lower gas prices.
* **“log(data$renewables\_generation\_ej)”**: This variable has a positive coefficient (0.28916) and a highly significant p-value (p < 0.001). It suggests that increased generation of renewable energy is associated with higher gas prices.
* **“log(data$nuclear\_energy\_generation\_twh)”**: The coefficient for this variable is positive (0.04442), but it has a relatively high p-value (0.341429). This p-value suggests that the relationship between nuclear energy generation and gas prices may not be statistically significant in this model.
* **“log(data$hydroelectricity\_generation\_twh)”**: This variable has a positive coefficient (0.14511) and a relatively high p-value (0.426892). The p-value suggests that the relationship between hydroelectricity generation and gas prices may not be statistically significant in this model.

In summary, based on this linear regression model, the variables that are statistically significant in explaining U.S. gas prices are oil production, oil crude oil prices, and renewables generation. These variables indicate that higher oil production and crude oil prices are associated with higher gas prices, while increased generation of renewable energy is also linked to higher gas prices. However, the relationship between gas prices and variables such as oil consumption, primary energy consumption per capita, nuclear energy generation, and hydroelectricity generation appears to be less statistically significant in this model.

#### LASSO

We then explored other feature selection algorithms that performs better than a simple linear regression, and ultimately settled on LASSO Regression. LASSO (Least Absolute Shrinkage and Selection Operator) regression offers several advantages over traditional linear regression for feature selection. It automatically selects relevant variables by shrinking some coefficients to zero, effectively eliminating irrelevant features from the model. LASSO handles multicollinearity by reducing coefficients of highly correlated variables, improving model stability. The sparsity induced by LASSO simplifies the model and enhances interpretability by highlighting the most important predictors. It also prevents overfitting by shrinking coefficients and promotes better generalization to unseen data. Furthermore, LASSO is flexible in handling large feature sets, making it suitable for high-dimensional datasets. These benefits make LASSO regression a powerful technique for feature selection, enhancing model interpretability, and improving predictive performance.

The results from the LASSO regression model on U.S. gas prices data reveal the estimated coefficients for the selected variables. The coefficients indicate the strength and direction of the relationships between the predictors and the gas prices.

* **(Intercept)**: The intercept term represents the estimated baseline gas price when all predictors are set to zero. In this case, the estimated intercept is 0.6026, suggesting that there is a baseline gas price even in the absence of any predictors.
* **US.GDP**: The coefficient for the US.GDP variable is 0.00005239. This positive coefficient suggests that an increase in the U.S. GDP is associated with a slight increase in gas prices. However, the small magnitude of the coefficient indicates a relatively weak relationship.
* **oil\_crude\_oil\_prices**: The coefficient for the oil\_crude\_oil\_prices variable is 0.02054. This positive coefficient indicates that higher crude oil prices are associated with higher gas prices. This suggests a positive relationship between crude oil prices and gas prices, which is expected since gas prices are influenced by the cost of crude oil.
* **primary\_energy\_consumption\_per\_capital**: The coefficient for the primary\_energy\_consumption\_per\_capital variable is -0.00223. This negative coefficient suggests that an increase in per capita primary energy consumption is associated with a slight decrease in gas prices. This relationship implies that higher energy consumption per capita might lead to more efficient energy use or a decrease in gas demand, resulting in lower gas prices.
* **coal\_consumption\_ej**: The coefficient for the coal\_consumption\_ej variable is 0.02388. This positive coefficient suggests that higher coal consumption is associated with higher gas prices. This indicates a positive relationship between coal consumption and gas prices, although the magnitude of the coefficient is relatively small.

Overall, these results provide insights into the determinants of U.S. gas prices. The positive coefficients for oil\_crude\_oil\_prices and coal\_consumption\_ej suggest that both crude oil prices and coal consumption have a positive influence on gas prices. On the other hand, the negative coefficient for primary\_energy\_consumption\_per\_capital suggests that per capita primary energy consumption has a slight negative association with gas prices. It’s important to note that the interpretation of these coefficients should consider the context of the dataset and the specific assumptions and limitations of the LASSO regression model.

### Prediction

This is the second part of the project where we aim to use avaiable data to make predictions on U.S. gas prices. Two machine learning algorithms, namely the neural network and random forest, are employed for this purpose. The utilization of both the neural network and random forest algorithms allows for a comprehensive exploration of different modeling approaches and increases the chances of developing accurate and reliable predictions for U.S. gas prices.

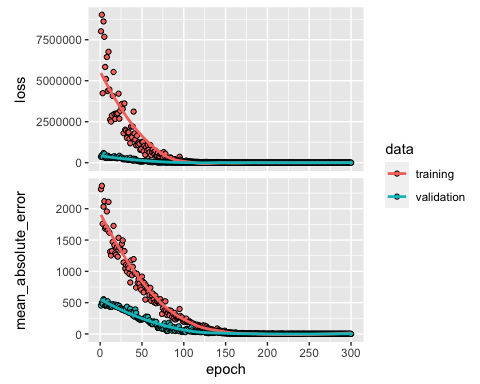
#### Neural Network

The firs algorithm implented is an one-layer Neural Network. The neural network algorithm, implemented using libraries such as TensorFlow and Keras, can capture complex non-linear relationships between the predictors and gas prices. By leveraging deep learning techniques, the neural network model can discover intricate patterns and dependencies in the data, enabling accurate predictions.

The dataset is partitioned into a training set and a test set, with the latter comprising one-third of the data, ensuring a robust evaluation of the model’s performance. The neural network architecture is defined with a hidden layer comprising 50 units, adopting the Rectified Linear Unit (ReLU) activation function. To mitigate overfitting, a dropout layer with a rate of 0.4 is incorporated, randomly setting 40% of the activations from the previous layer to zero during each iteration of stochastic gradient descent (SGD). The output layer consists of a single unit, capturing the desired prediction. For optimizing the model’s parameters, the RMSprop optimizer is employed, dynamically determining the step size during gradient descent. The mean squared error (MSE) loss function is utilized to evaluate the model’s performance, with the mean absolute error serving as an additional metric during both training and testing. The model is trained using the training data, employing a batch size of 32 and 300 epochs, each epoch involving approximately 5.5 SGD steps. To assess the model’s progress and fine-tune its performance, the validation data, representing the test set, is employed, and the corresponding results are stored in the “history” variable, allowing for comprehensive analysis and evaluation.

#### Visual Representation of Neural Network Performance

plot(history)



The plot represents the performance of a neural network model in predicting U.S. gas prices over multiple epochs. The decreasing loss values indicate that the model quickly learns and improves its predictive capabilities during the initial phase of training. This suggests that the model successfully captures the underlying patterns and relationships in the data related to U.S. gas prices. The exponential decrease in the loss signifies that the model effectively adapts to the training data and makes accurate predictions.

As the training progresses, the loss curve gradually converges towards zero, indicating that the model’s performance continues to improve, albeit at a slower pace. This convergence suggests that the neural network model is approaching its optimal predictive capacity for U.S. gas prices. The consistent and relatively low val\_loss values throughout the training process indicate that the model generalizes well to unseen data, suggesting that it can effectively predict U.S. gas prices beyond the training dataset. The final

In summary, the plot demonstrates that the neural network model is highly effective in predicting U.S. gas prices. The significant decrease in the loss and the convergence towards zero showcase the model’s ability to capture the complex relationships between various predictors and gas prices. The consistency between the loss and val\_loss curves highlights the model’s robustness and its capability to provide accurate predictions for U.S. gas prices, enabling valuable insights for decision-making and analysis in the energy sector.

#### Random Forrest/Decision Trees

The second algorithm used is a tree model, evovling from a simple decision tree to a random forest. Random forest algorithm operates by constructing an ensemble of decision trees and making predictions based on their collective outputs. This algorithm is adept at handling high-dimensional datasets and can effectively capture interactions between predictors. By leveraging the diversity of decision trees, random forest models provide robust predictions and can handle noisy or missing data. evolving from a simple decision tree to Random Forrest.

For this project, it begins by building a bagged regression forest with default parameters, resulting in a mean debiased error of 0.1348957. The code then constructs another bagged regression forest with 5000 trees and mtry = 24, yielding a slightly higher mean debiased error of 0.135718. Additionally, a random forest with 5000 trees and mtry = 5 is built, producing a mean debiased error of 0.1606098. The code further includes making predictions and plotting confidence intervals based on the bagged regression forest and random forest models. Specifically, it predicts gas prices for the 30th observation and stores the predictions, standard errors, and the true gas price in the pred.data dataframe.

#### Visual Representation of Performance for Different Tree Models

# Plot confidence intervals  
pd <- position\_dodge(0.8)  
ggplot(pred.data, aes(x = type, y = pred, group = type)) +  
 geom\_point(position = pd) +  
 geom\_errorbar(data = pred.data,  
 aes(ymin = pred - 1.96 \* se, ymax = pred + 1.96 \* se, color = type),  
 width = 0.1) +  
 geom\_hline(yintercept = y[idx], linetype = "dashed", color = "black") +  
 geom\_text(aes(0, y[idx], label = "Truth", vjust = -1, hjust = -0.5))

