Fuelonomics: Analyzing Gas Price Dynamics through Machine Learning

Yimo Shen

2023-06-04

Introduction

Gas prices in the U.S. are of significant concern to individuals and society as a whole due to their direct impact on personal finances, transportation costs, and the overall economy. People care about gas prices because they affect their daily lives and budgeting decisions. Fluctuations in gas prices can significantly influence the affordability and accessibility of transportation, impacting individuals' ability to commute to work, run errands, or travel long distances. Additionally, gas prices have broader implications for the economy, as they can affect the cost of goods and services due to transportation costs being passed onto consumers. Understanding the factors that cause gas prices to fluctuate is crucial as it enables policymakers, economists, and consumers to make informed decisions. By identifying the underlying drivers of price changes, policymakers can implement strategies to mitigate volatility, ensure energy security, and promote a more sustainable and efficient transportation system. Moreover, consumers can plan their budgets and make informed choices about vehicle purchases, fuel efficiency, and alternative energy sources, contributing to environmental sustainability and financial stability.

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

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$), suggesting 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$), indicating 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), suggesting 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 (*), indicating 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$), suggesting 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), indicating 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), suggesting 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.

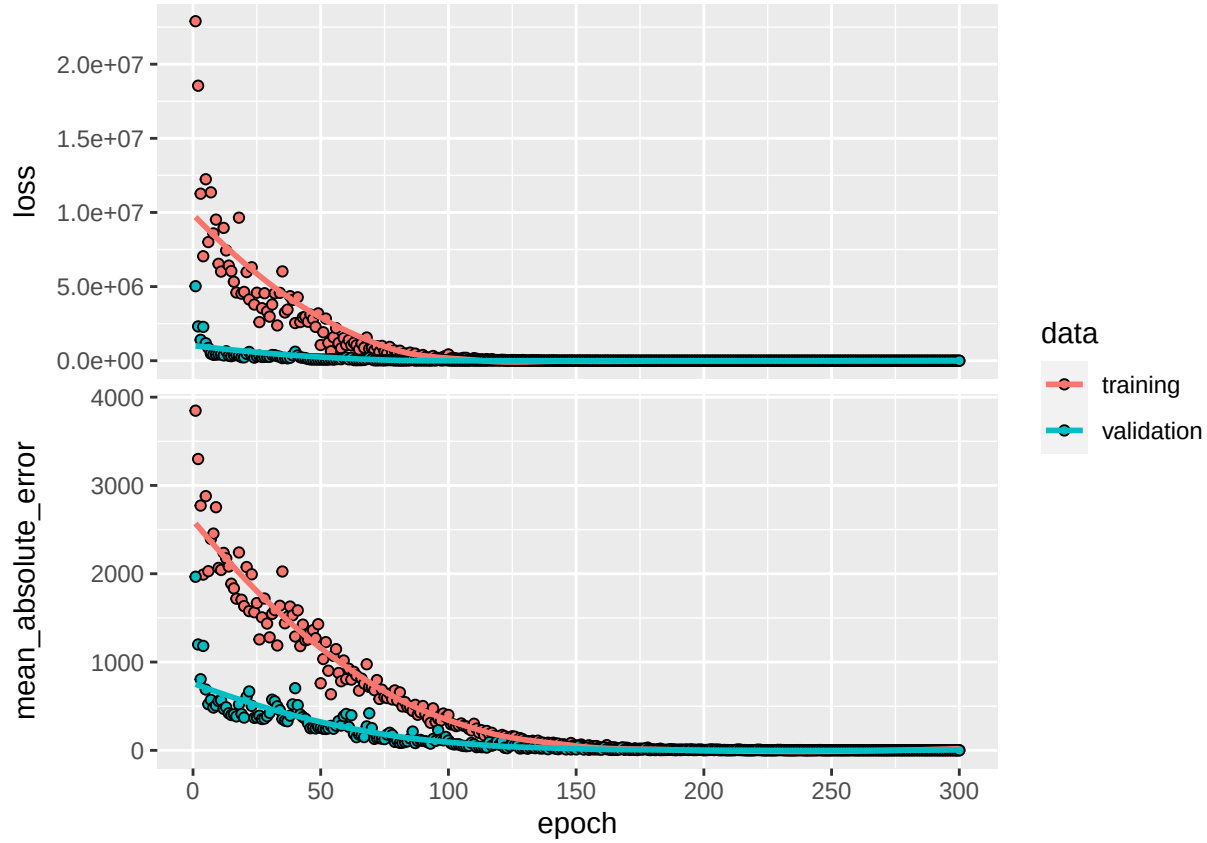
Overall, these results provide insights into the determinants of U.S. gas prices. The positive coefficients for oil_crude_oil_prices and coal_consumption_ej suggest that both crude oil prices and coal consumption have a positive influence on gas prices. On the other hand, the negative coefficient for 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

Neural Network The first prediction algorithm implemented 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 is employed, and the corresponding results are stored in the "history" variable, allowing for comprehensive analysis and evaluation. The final model's test error was 10.8%.

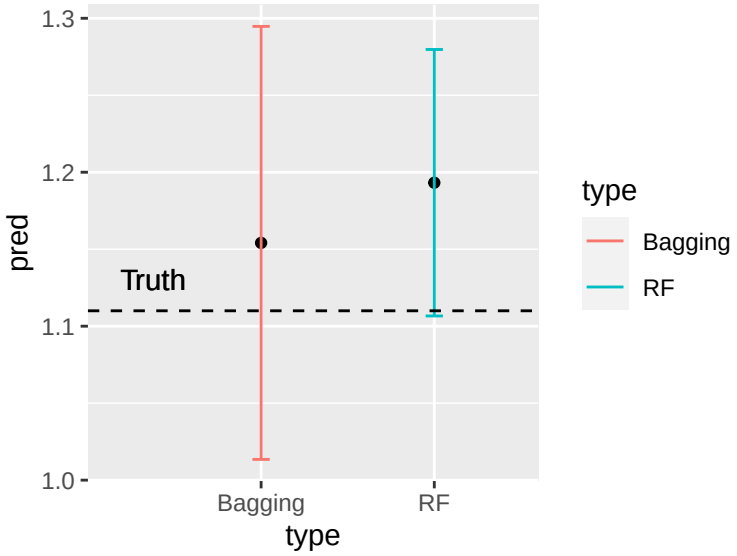
Visual Representation of Neural Network Performance



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.

Random Forest We have implemented three iterations of tree methods, starting from a simple decision tree and progressing to a random forest with bagging. The decision tree method constructs a tree-like model by recursively splitting the dataset based on various features to make predictions, offering interpretability but suffering from overfitting. To address this, the code advances to the random forest algorithm, which combines multiple decision trees through bagging. Bagging involves training each tree on a random subset of the data, reducing overfitting risk and providing more accurate and stable results. Random forests capture non-linear relationships, handle high-dimensional data, and effectively handle missing values or outliers. The code's progression from a decision tree to a random forest with bagging aims to improve prediction accuracy and generalizability, leveraging the strengths of individual trees while mitigating their weaknesses. By utilizing this ensemble approach, the code endeavors to provide a more robust and reliable model for forecasting gasoline prices in the US.

Visual Representation of Performance for Random Forest



This visual focuses on comparing the confidence intervals of bagging and random forest models in predicting gas prices in the U.S. The code starts by defining the position dodge (pd) for the side-by-side visualization of the intervals. Using the ggplot library, a plot is created, with the x-axis representing the model type (bagging or random forest) and the y-axis representing the predicted gas prices. The plot includes points representing the predicted values for each model, and error bars are added to represent the confidence intervals. The error bars are constructed based on the standard errors (se) of the predictions, calculated as 1.96 times the standard error, representing a 95% confidence interval. Additionally, a dashed horizontal line is drawn at the true gas price (y[idx]) for reference. Finally, a text label "Truth" is added to the plot, positioned at (0, y[idx]), highlighting the actual gas price. Overall, this visual provides a comparison of the confidence intervals between bagging and random forest models, suggesting that from a performance standpoint, the bagging algorithm is actually more effective in predictions, aiding in understanding the uncertainty associated with the predicted gas prices.

Conclusion

In conclusion, our project successfully address the original question of what are the factors that affect gasoline prices in the U.S. The findings obtained using feature selection models like LASSO regression and linear regression reveals important factors influencing gas prices: crude oil prices, oil consumption, renewable energy generation and coal consumption have a positive impact on gas prices, while factors like per capita primary energy consumption has a slight negative association with gas prices . It is crucial to consider the dataset's context, as well as the assumptions and limitations specific to the LASSO and linear regression model when interpreting the results. Furthermore, our investigation into prediction demonstrated that across all algorithms employed, the mean square error was remarkably small. This finding suggests that the limitations imposed by the dataset, which only covers the period from 1965 to 2023 and lacks more granular temporal information, pose challenges in making accurate predictions. Therefore, it is important to acknowledge the limitations inherent in the dataset when interpreting the predictive performance of the models.

Overall, this project shed light on the determinants of gas prices in the U.S., providing valuable insights for understanding the factors influencing this crucial aspect of the energy market. However, the findings should be interpreted cautiously, considering the context of the dataset, specific modeling assumptions, and limitations. Future research with more comprehensive and detailed data could further enhance our understanding of gas price determinants and improve prediction accuracy.

APPENDIX

CODE

```
# data <- read.csv("gas_prices_data.csv")
# sum(is.na(data$`US Gasoline Prices Adjusted for Inflation`))
# data <- na.omit(data)  # remove all rows with NAs
# #dim(data)
# #data <- column_to_rownames(data,'...1')
# data <- data[-1, -1]
# data <- as.data.frame(apply(data, 2, as.numeric))  # Convert all variable types to numeric
# sapply(data, class)
# data <- data[, -(6:9)]
```

DATA PREP

```
# x <- model.matrix(`US.Gasoline.Prices`~., data)[, -1]
# y <- data$US.Gasoline.Prices
```

SPLIT DATA

LINEAR REGRESSION

```
# set.seed(123) # Split data into training & test sets to estimate the test error
# grid <- 10^seq(10, -2, length=100)
# train <- sample(1:nrow(x), 0.5*nrow(x), replace = FALSE)  # half training, half test
# x.test <- x[-train,]
# y.test <- y[-train]
#
# cv.out.lasso <- cv.glmnet(x[train,], y[train], alpha=1)
# plot(cv.out.lasso)
# bestlam.lasso <- cv.out.lasso$lambda.min #the best lambda is 0.01197346
#
# lasso.mod <- glmnet(x[train,], y[train], alpha=1, lambda=grid)
# lasso.pred <- predict(lasso.mod, s=bestlam.lasso, newx = x.test)
# mean((lasso.pred - y.test)^2)# test MSE is 0.009198251
#
# lasso.out <- glmnet(x, y, alpha=1)# whole data, default grid
# lasso.pred.out <- predict(lasso.out, type="coefficients", s=bestlam.lasso)[1:19, ]
# #lasso.pred.out  # a few coefficients are exactly zero
# selected_features_lasso <- lasso.pred.out[lasso.pred.out!=0]
# selected_features_lasso
```

LASSO

```
# #### ------ NN ------
# ### -- Load & format the gas prices data --
# set.seed(123)
# ntest <- floor(nrow(data)/3)  # 1/3 of data is test
# testid <- sample(1:nrow(data), ntest)  # indices of test obs
#
# ## Set up model structure: NN with 1 hidden layer of 50 units
# # Note: the pipe operator %>% passes the previous term as the first argument to the next function, an
# modnn <- keras_model_sequential() %>%
#   layer_dense(units=50, activation="relu", input_shape=ncol(x)) %>%  # hidden layer
#   layer_dropout(rate=0.4) %>%  # dropout layer (regularization), in which a random 40% of the 50 acti
#   layer_dense(units=1)  # output layer
#
# modnn %>% compile(loss="mse", optimizer=optimizer_rmsprop(),
#                   metrics=list("mean_absolute_error"))  # in addition to MSE, also keep track of trai
#
# ## Supply the training data & 2 parameters, batch size=32 and epochs=300, where 1 epoch is 176/32=5.5
# history <- modnn %>%
#   fit(x[-testid,], y[-testid], epochs=300, batch_size=32,
#       validation_data=list(x[testid,], y[testid]))  # keep track of progress
```

Neural Network

```
# plot(history)
```

Visual Representation of Neural Network Performance

```
# # Build a bagged regression forest using default parameters
# x <- data[, -1]
# y <- data$US.Gasoline.Prices
# bag.data <- regression_forest(x, y, mtry = 24)
# mean(bag.data$debiased.error)
# # 0.1348957
#
# # Build a bagged regression forest with 5000 trees and mtry = 24
# bag.data <- regression_forest(x, y, num.trees = 5000, mtry = 24)
# mean(bag.data$debiased.error)
# # 0.135718
#
# # Build a random forest with 5000 trees and mtry = 5
# rf.data <- regression_forest(x, y, num.trees = 5000, mtry = 5)
# mean(rf.data$debiased.error)
# # 0.1606098
#
# # make predictions and plot confidence intervals
```

```
# idx <- 30
# bag.pred <- predict(bag.data, x[idx, ], estimate.variance = TRUE)
# rf.pred <- predict(rf.data, x[idx, ], estimate.variance = TRUE)
# y[idx]
# pred.data <- data.frame("type" = c("Bagging", "RF"), pred = c(bag.pred$predictions, rf.pred$predictio
```

Tree Models, Random Forest

```
# # 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))
```

Visual Representation of Performance for Different Tree Models

SOURCES

- BP. (n.d.). Statistical Review of World Energy. Retrieved from https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html
- U.S. Department of Energy. (2012, August 20). Fact #741: August 20, 2012 - Historical Gasoline Prices, 1929-2011. Retrieved from https://www.energy.gov/eere/vehicles/fact-741-august-20-2012-historical-gasoline-prices-1929-2011
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: with Applications in R (2nd ed.). Springer.
- Tatić, M. (2020, February 12). Power of XGBoost-LSTM in Forecasting Natural Gas Price. Towards Data Science. Retrieved from https://towardsdatascience.com/power-of-xgboost-lstm-in-forecasting-natural-gas-price-f426fada80f0