

Data Fusion and Model-based Integration For Retail Gasoline Price Prediction

2025

Matt Collins, Garrett Gall, Ethan Vogelman, Andrew Mason, Zach Brown

1. Abstract

This project demonstrates how model-based data fusion techniques improve the prediction of U.S. retail gasoline prices.

Three heterogeneous data views: crude oil market indicators, macroeconomic conditions, and region-specific supply/demand metrics, were independently preprocessed and aligned to a common temporal structure. Linear Regression w/ LASSO, Random Forest, and XGBOOST were trained on each view. Predictive performance for the individual models was evaluated using RMSE, MAE, and MAPE, revealing complementary information across the distinct feature classes. To take advantage of the complementary signals, a weighted fusion approach was applied to combine the view-specific predictions into a single ensemble estimate.

The final fused model achieved an RMSE of 0.0630, MAE of 0.2259, and MAPE of 8.25%, indicating an average deviation of approximately 6.3 cents or 8.25% from true gasoline prices. The results show that integrating multiple data sources meaningfully improves the model's ability to capture long-term pricing dynamics as opposed to relying on a single view.

Perfect forecasting of gas prices is unrealistic because the driving factors are multivariate (driven by many systems), nonlinear (endless feature interactions), and sensitive to unpredictable and likely unobserved market events. The complexity of this task is exactly why a data fusion strategy is valuable: no single data view captures all drivers, but the fusion of multiple complementary views can significantly improve predictive performance.

2. Background

Gasoline prices fluctuate in response to a diverse set of global, domestic, and regional factors. To explore whether near-future gas prices can be forecasted with reasonable accuracy, this project asks:

Research Question: Can we develop a predictive model capable of forecasting weekly gasoline prices for a given region with usable accuracy?

To approach this systematically, gasoline price formation was decomposed into three conceptual layers of feature sets:

2.1 Crude Oil Market View

This view captures the supply-side fundamentals that form the baseline cost of gasoline. Features included:

- WTI, Brent, and RBOB spot prices
- Crack spreads (refining margins)
- WTI–Brent differentials
- Stock returns of major integrated oil companies (proxy for market expectations)

2.2 Macroeconomic View

Macroeconomic indicators influence demand conditions and purchasing power. This view included:

- Federal Funds Rate
- Interest Rates
- Dollar Index
- Consumer Price Index (CPI)
- Unemployment Rate

These factors were included to attempt to capture inflationary pressures, consumption strength, and broad economic conditions that may influence the price at the pump.

2.3 Regional Market View

Regional variables explain why local pump prices vary relative to the national average. Regional predictors included:

- Crude oil production
- Gasoline stocks
- Gasoline imports
- Finished motor gasoline production & supply
- Crude oil imports
- Refinery utilization rates

Together, these views support the development of a 3-layer predictive model that integrates information from the global oil market, domestic macroeconomic conditions, and regional supply conditions.

2.4 EIA PADD Regional System

Much of the data used in model training was gathered from the U.S. Energy Information Administration (EIA). The EIA organizes its petroleum, refining, and gasoline data using the Petroleum Administration for Defense Districts (PADD) system. It is the standard geographic framework used by federal agencies, refineries, pipelines, and commodity analysts to report and analyze petroleum supply and demand.

The EIA divides the United States into five primary PADD regions, each representing a distinct petroleum market with unique refinery capacity, supply chain configurations, and consumption patterns. For the task of gasoline price prediction, these regional distinctions are critical because they capture structural heterogeneity in supply, production, imports, and regional disruptions.

3. Methodology

Evaluation of the Fusion Approach

The multi-view weighted averaging method combines heterogeneous information sources into a unified prediction.

This data integration strategy aligns with real-world price formation processes, where:

- **Oil market movements:** drive baseline national pricing
- **Macroeconomic conditions:** shift demand and cost pressures
- **Regional factors:** modulate localized price deviations

The weighted fusion model balanced these influences while giving more weight to views that produced models with better evaluation metrics on the unseen validation set.

Data Preprocessing/Preparation

Each view was standardized to a weekly Sunday index spanning **January 1, 2006 – May 8, 2022**, chosen to maximize overlapping coverage across crude, regional, and macroeconomic sources. Data processing involved:

- Aligning date ranges
- Ensuring consistent weekly indexing

- Handling missing observations
- Standardization of predictor features for Lasso Regression

Each view was split:

- **X** (Predictors)
- **Y** (Target: U.S. regular gas price)

A **time-series split** for 5-fold cross-validation was used to preserve chronological order during training and evaluation.

Each dataset/view was trained and evaluated with the following ML algorithms:

- XGBoost
- Random Forest
- Lasso Regression

Model Selection:

1. XGBoost (Extreme Gradient Boosting)

XGBoost was selected because it is one of the most powerful and widely used algorithms for when predictors interact in nonlinear and hierarchical ways.

Why XGBoost?:

- Captures complex nonlinear relationships: Gasoline prices depend on interactions between macromarket features- relationships that are rarely linear.
- Handles collinearity well: Variables move together, and boosting methods are good at managing correlated features.
- Built-in regularization: Prevents overfitting while still allowing flexibility.
- High predictive accuracy: Tends to outperform simpler models on economic and energy datasets.

2. Random Forest

Random Forest provides a second nonlinear model but uses bagging instead of boosting. Allows us to test how the model structure (boosting vs. bagging) affects performance.

Why Random Forest?:

- Robust to noise and outliers: The gas market can be very noisy and volatile during shocks (supply chain/economic disruptions), and Random Forest is known to be stable under noisy conditions
- Captures nonlinearities and interactions: Useful for modeling supply chain behavior where multiple regional variables jointly influence prices.
- Less risk of overfitting than single decision trees: Averaging reduces variance.
- Interpretability through feature importance and SHAP: Helps with the task of explaining regional drivers, not just predicting prices.

3. Lasso Regression

Linear Regression w/ Lasso Regularization was selected to provide a linear and highly interpretable model with built-in feature selection through its penalty term, shrinking coefficients down to zero.

Why Lasso?:

- Acts as a linear benchmark: Helps determine whether nonlinear models (XGBoost and RF) meaningfully improve performance.
- Improved interpretability: Coefficients directly measure the feature relationship to the predictor.

Evaluation Metric Selection:

We evaluated model performance using RMSE, MAE, and MAPE to capture different aspects of predictive accuracy and to ensure robustness across multiple regions and error distributions.

1. RMSE (Root Mean Squared Error)

RMSE penalizes larger errors more heavily due to the squaring term.

It was selected because:

- Standard benchmark metric for regression models
- Captures performance during volatile periods, not just stable ones
- Large deviations matter in the energy market

2. MAE (Mean Absolute Error)

MAE treats all errors equally and is easier to interpret than RMSE.

Chosen because:

- Robust to outliers, provides a balanced view of model error
- Provides a stable, intuitive measure of typical model deviation.
- Complements RMSE by reducing the impact of extreme events.

3. MAPE (Mean Absolute Percentage Error)

MAPE expresses error as a percentage, which is often more interpretable and allows comparison across regions with different price levels.

Chosen because:

- Gasoline prices vary across Regions; percentage errors normalize comparisons.
- Easy to interpret: (Predicted gas price was off by 8.7%).
- Explains where the model struggles proportionally, not just in magnitude.

Model 1: Crude Oil

The crude oil prices dataset was divided twice; once to separate X and y columns, and once to apply a time series split to both tibbles. The time series was set to apply 5-fold cross-validation to the sets, ensuring that the training and test sets were split chronologically rather than randomly, and that the timestamps in the training data came before those in the test data.

The X set, “X_crude,” contained the following columns:

- Weekly stock price indicators:
 - Return_Stock_COP
 - Return_Stock_CVX
 - Return_Stock_MRO
 - Return_Stock_SLB
 - Return_Stock_VLO
 - Return_Stock_XOM
- Weekly oil price indicators:
 - Brent_Close
 - WTI_Close
 - RBOB_Close
- CrackSpread: Difference between the price of crude oil and its refined products
- WTI_Brent_Spread: Difference between the price of WTI and Brent crude oils

The y set, “y_crude,” contained only the GasPrice_US column, which denotes weekly U.S. national regular gas prices.

Further steps were taken to properly scale the data for Lasso Regression via scikit-learn StandardScaler for both the X and y tibbles. Data was then fit to three models for training.

The following hyperparameters were tuned for training XGBoost:

- 500 estimators
- Learning rate of 0.05
- Max depth of 6
- Subsample 0.8
- Histogram tree method (faster for larger datasets)

The following hyperparameters were tuned for training the Random Forest:

- 400 estimators
- Unlimited max depth

The following hyperparameters were tuned for training Lasso Regression:

- Alpha 0.01
- 10,000 max iterations

Model predictions were evaluated against true prices with Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE).

Model 2: Regional Market Indicator Training

The regional indicator dataset needed to be one-hot encoded from categorical entries to dummy values before being fit to a model. This was done via the `get_dummies` function in pandas, which assigns binary values to categorical variables to make them viable for machine learning. The regional indicator dataset was then divided twice in the same manner as the crude oil indicator dataset, with an X and y split, then a 5-fold cross-validated time series split. The X dataset, “`X_region`,” contained the following columns:

- Region (EIA PADD geographic regions)
- Total Gasoline Stocks
- Total Gasoline Imports
- Motor Gasoline Production
- Crude Oil Imports
- Refinery Utilization Rate

The y dataset, “`y_region`,” contained the Gasoline Price column, similar to the structure of the previous Crude Oil Indicator datasets.

Scaling for Lasso Regression and hyperparameter tuning remained the same as in the previous Crude Oil Indicator model training.

All models were evaluated with the same metrics - RMSE, MAPE, and MAE.

Macroeconomic Indicator Training

The X set, “X_macro,” contained weekly updates of several macroeconomic factors in the United States in the following columns:

- fed_funds_rate
- interest_rate
- dollar_index
- cpi (Consumer Price Index)
- unemployment_rate

The y dataset, “y_macro,” contained the gas_price column, similarly structured to the previous two sets.

Datasets were scaled and split identically to the previous Crude Oil and Regional Indicator datasets. Hyperparameters remained identical for training.

Performance was evaluated using the same metrics: RMSE, MAPE, and MAE.

Feature Importance

Feature importance was calculated for the results from the trained Lasso Regression model by taking the weight/coefficient of each feature in the regularized regression formula. Feature importance charts were produced to analyze and compare feature importance.

3-Model Fusion with Weighted Average

The approach to data integration was to merge predictions from multiple models on heterogeneous views. This was achieved via the weighted average from the best-performing models from each data view: Crude Oil, Regional, and Macroeconomic Indicators.

In this particular study, Lasso Regression consistently outperformed XGBoost and Random Forest in RMSE, MAPE, and MAE performance metrics, so it was chosen across the board to be used for the weighted average integration. Using the performance metrics gained from model testing, weights were calculated with the following formula:

$$w_i = \frac{\frac{1}{metric_i}}{\sum_j (\frac{1}{metric_j})}$$

The final weighted averages for our metrics, completing our integration, were then calculated using the following formula, where the weights add up to 1:

$$avg_{metric} = (w_{crude} * metric_{crude}) + (w_{regional} * metric_{regional}) + (w_{macro} * metric_{macro})$$

4. Results

To evaluate whether gasoline prices can be reliably predicted using a combination of crude oil, regional, and macroeconomic factors, we applied three predictive modelling approaches—XGBoost, Random Forest, and LASSO Regression—using 5-fold time-series cross-validation. In this section, we present the resulting metrics, chosen before model building: root mean squared error (RMSE), mean absolute percentage error (MAPE), and mean absolute error (MAE). We also display the feature importances of each LASSO regression model, as it achieved the best performance for each view. Finally, we present the final metrics for evaluating the approach’s accuracy, calculated using the weighted average integration equation outlined in the Methodology section.

Crude Oil Results

First, the crude oil view produced the following evaluation results for each model type, after averaging over the five folds:

XGBoost:

Average RMSE: 0.0516

Average MAPE: 0.0669

Average MAE: 0.1781

Random Forest:

Average RMSE: 0.0470

Average MAPE: 0.0635

Average MAE: 0.1688

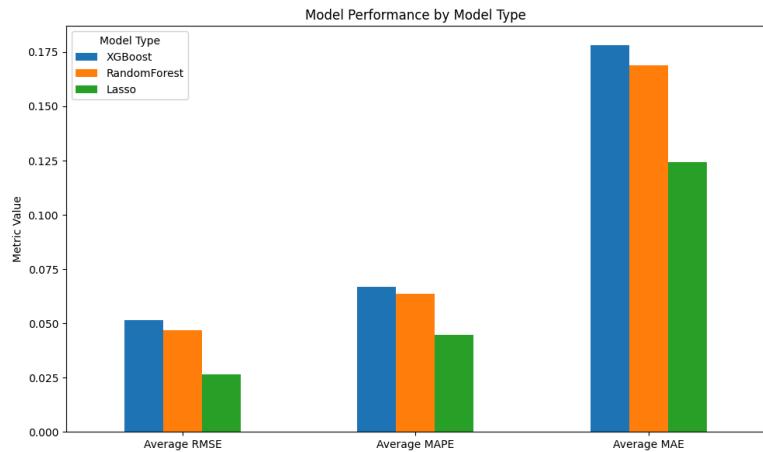
LASSO Regression:

Average RMSE: 0.0264

Average MAPE: 0.0445

Average MAE: 0.1243

Figure 4.1



While the XGBoost and Random Forest models produced extremely similar results for all metrics, the LASSO Regression model, denoted by the color green in Figure 4.1, performed much better in all metrics.

The importance of each feature in the LASSO Regression model, determined by the absolute value of the coefficients, is as follows, from most to least:

Table 4.1

RBOB_Close	0.385493
Brent_Close	0.157868
Return_Stock_SLB	0.019461
Return_Stock_MRO	0.015958
WTI_Brent_Spread	0.011699
Return_Stock_VLO	0.004941
Return_Stock_COP	0.001030
Return_Stock_CVX	0.000000
Return_Stock_XOM	0.000000
WTI_Close	0.000000
CrackSpread	0.000000

Table 4.1 reveals that the features RBOB_Close and Brent_Close are significantly more important to the model's predictions than the others.

Regional Results

The regional data view produced the following evaluation results for each model type, after averaging over the five folds:

XGBoost:

Average RMSE: 0.5112

Average MAPE: 0.2246

Average MAE: 0.5807

Random Forest:

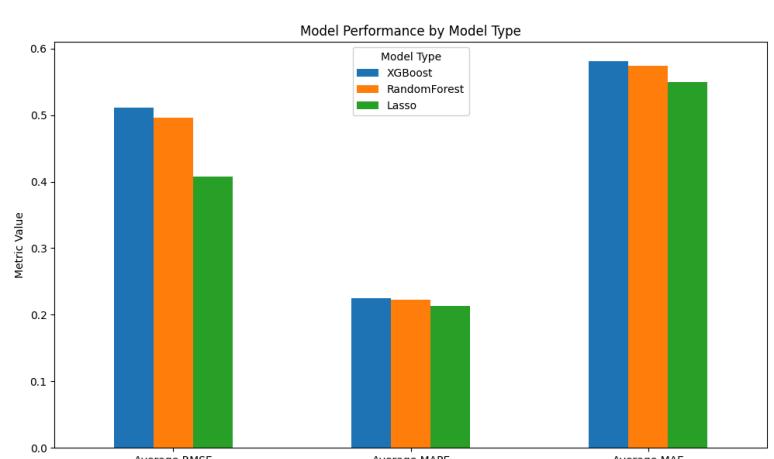
Average RMSE: 0.4958

Average MAPE: 0.2225

Average MAE: 0.5743

LASSO Regression:

Figure 4.2



Average RMSE: 0.4071

Average MAPE: 0.2130

Average MAE: 0.5491

Similar to the crude oil view, the LASSO Regression model ended with the best performance of the three types in all metrics, although the gap between models was narrower, as shown in Figure 4.2.

The importance of each regional feature in the LASSO Regression model, determined by the same process as before, is displayed below from most to least:

Table 4.2

Total Gasoline Stocks	0.262877
PADD 4*	0.184471
PADD 3*	0.027604
Refinery Utilization Rate	0.017151
PADD 1*	0.013129
PADD 2*	0.011959
PADD 5*	0.000000
Crude Oil Inputs	0.000000
Motor Gasoline Production	0.000000
Total Gasoline Imports	0.000000

Table 4.2 reveals that the coefficients of ‘Total Gasoline Stocks’ and ‘PADD 4’ have the greatest contribution to the model’s predictions, while regularization shrunk the coefficients of other variables to zero.

Macroeconomic Results

The macroeconomic data view produced the following evaluation results for each model type, after averaging over the five folds:

XGBoost:

Average RMSE: 0.3323
Average MAPE: 0.1673
Average MAE: 0.4442

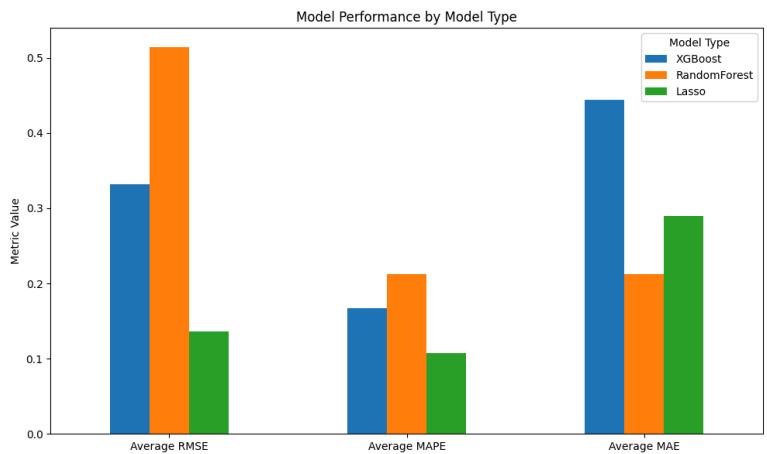
Random Forest:

Average RMSE: 0.5138
Average MAPE: 0.2125
Average MAE: 0.2125

LASSO Regression:

Average RMSE: 0.1358
Average MAPE: 0.1077
Average MAE: 0.2898

Figure 4.3



While the LASSO Regression model again performed best in the RMSE and MAPE metrics, the Random Forest actually had the lowest MAE. However, the LASSO model was decided upon as the best overall model.

The importance of each macroeconomic feature in the LASSO Regression model, determined once again using the coefficients, is displayed below from greatest to least:

Table 4.3

dollar_index	0.787172
cpi	0.584934
unemployment_rate	0.073732
fed_funds_rate	0.068314
interest_rate	0.000000

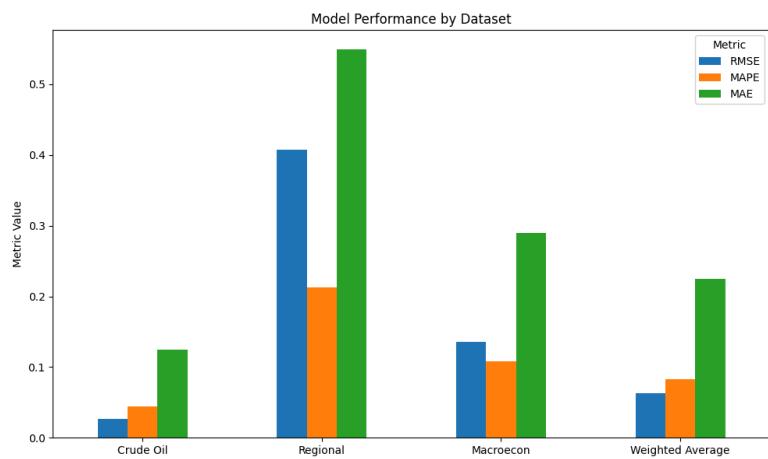
Table 4.3 reveals that the features dollar_index and cpi are significantly more important to the model's predictions than the others.

Final Weighted Average Results

After calculating a weighted average of the RMSEs, MAPEs, and MAEs of the best-performing model for each data view, in our case, the LASSO Regression model for all three, we produced the following final evaluation metrics:

RMSE: 0.0629
 MAPE: 0.0823
 MAE: 0.2253

Figure 4.4



Interpretation of Metrics

The weighted average RMSE of 0.0630 tells us that, on average, the predicted gas price is off by **6.3 cents** from the actual observed price.

The weighted average MAPE of 0.0825 tells us that, on average, the predicted gas price is within **8.25%** of the actual observed price.

The weighted average MAE of 0.2259 tells us that, on average, the predicted gas price is off by about **22.6 cents** from the actual observed price.

While RMSE and MAE have the same interpretation, RMSE penalizes outliers more heavily.

5. Discussion

The results of this study highlight several important insights about gasoline price prediction and multi-view data integration.

Predictive Value of the Three Data Views

Crude Oil Indicators emerged as the most stable and predictive view. This insight is unsurprising, considering that crude oil prices are the fundamental input into gasoline production and tend to exhibit smoother, economically driven patterns. The strong performance of Lasso

suggests that the dominant linear relationship between crude oil and retail gasoline price dismisses the need for nonlinear tree-based modeling.

Macroeconomic Indicators demonstrated moderate predictive power. Factors such as inflation, unemployment, and interest rates influence gasoline demand and macro-level purchasing capacity, but these relationships evolve gradually. Macroeconomic data alone is insufficient for predictive accuracy for our high-frequency, volatile target, but still contributes valuable context to the model.

Regional Indicators, while conceptually important, proved highly volatile and difficult to model. Regional refinery outages, weather disruptions, policy changes, and distribution bottlenecks introduce irregular fluctuations that are not fully captured by the available features. The relatively poor performance metrics in this layer point to the challenges of modeling local supply conditions without full-scope, high-resolution operational data.

SHAP Analysis of Regional Supply Chains

To interpret the predictive model and validate that it captures meaningful economic structure, we applied SHAP and produced its plots to quantify and visualize the contribution of the explanatory regional features to regional gasoline price deviations in Random Forest. Across all regions, SHAP consistently identified gasoline stocks, refinery utilization, and crude oil production as the dominant drivers of deviations. Interestingly enough, the largest contributions often come from other regions, not the region itself. This suggests that regional gasoline markets respond strongly to external supply chain conditions. We can infer that this is because the U.S. gasoline supply chain is highly interconnected through pipelines, shipping routes, and refinery networks.

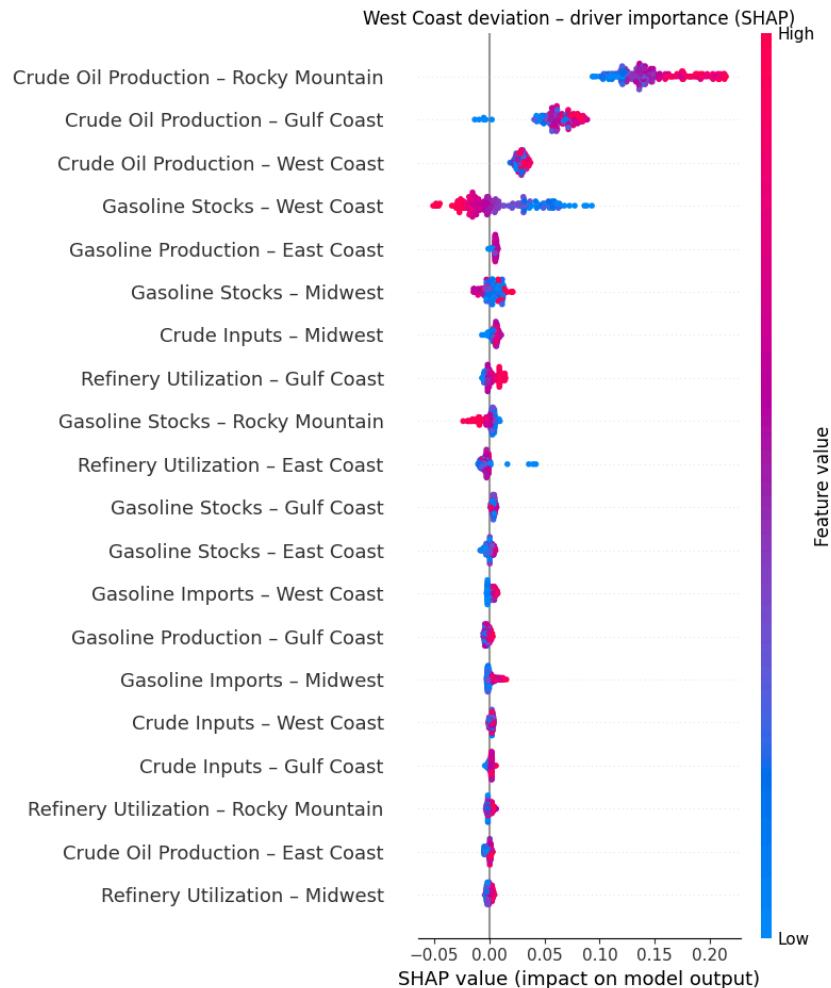


Figure 5.1- SHAP plot of Random Forest Regional Features for West Coast Price Deviation

The most interesting results come from the West Coast region. No other region reported such large feature impacts on SHAP magnitudes (0.15–0.25). The 3 most significant features being Crude Oil Production (Rocky Mountain, Gulf Coast, West Coast). We believe that this is the result of the West Coast region having limited refining capacity, so they depend heavily on production from the Rocky Mountain and Gulf Coast regions. This assumption is supported by the results of “Gasoline Stocks - West Coast,” whose SHAP distribution illustrates that low internal supply causes massive price volatility, or a huge deviation from the national average. The region’s price deviation being so sensitive to internal supply and production from external regions, point to supply chain isolation and limited refining capacity of the West Coast PADD region.

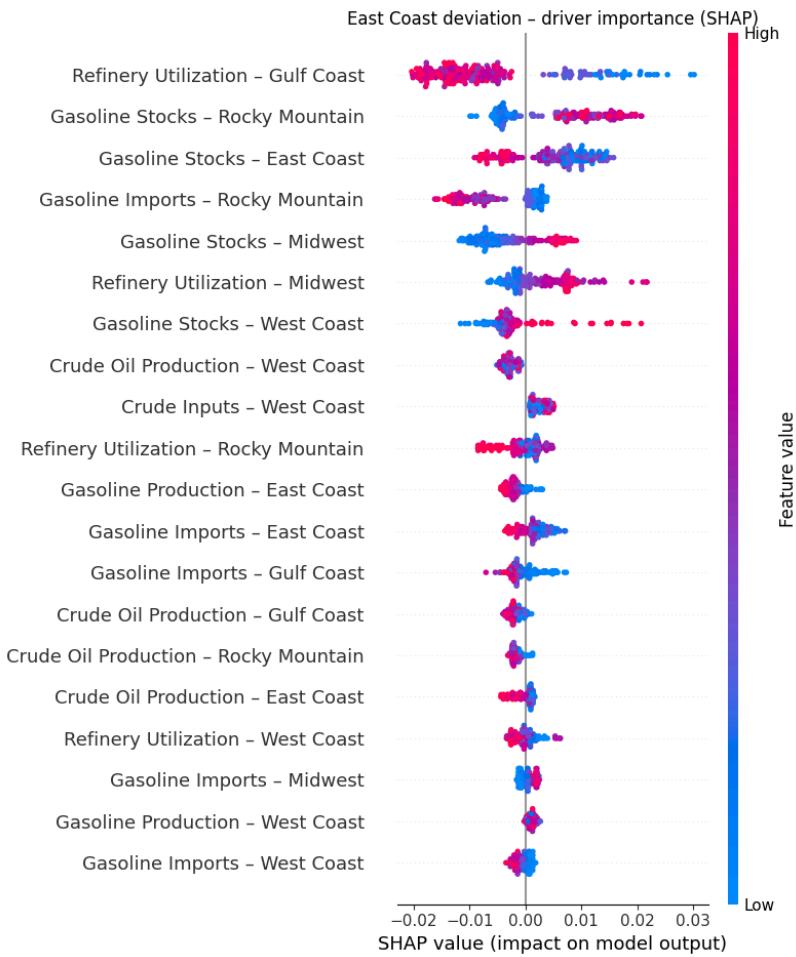


Figure 5.2- SHAP plot of Random Forest Regional Features for East Coast Price Deviation

East Coast deviation was heavily influenced by Gulf Coast refinery utilization and nationwide inventory, reflecting the region's heavy reliance on imported fuel and a limited refining capacity.

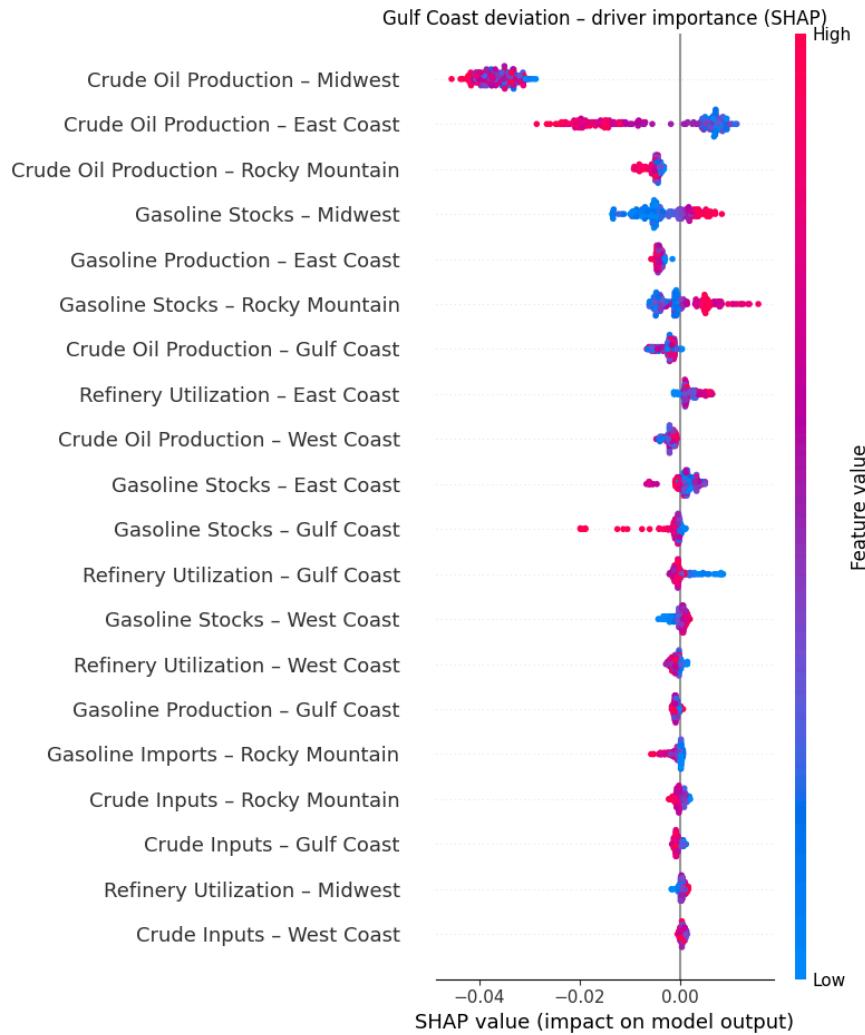


Figure 5.3- SHAP plot of Random Forest Regional Features for Gulf Coast Price Deviation

Gulf Coast price deviation is shaped by the supply and production in other regions. Suggests the region serves as the primary refining hub that balances supply for the entire country.

The SHAP results provide justification for the decision to construct region-specific data views and predictive models. The SHAP plots reveal that variables influencing price deviations differ substantially across regions. This directly supports the idea that a single unified model or data view would fail to capture the supply chain infrastructural dynamics of each region as effectively as tailored models for each region.

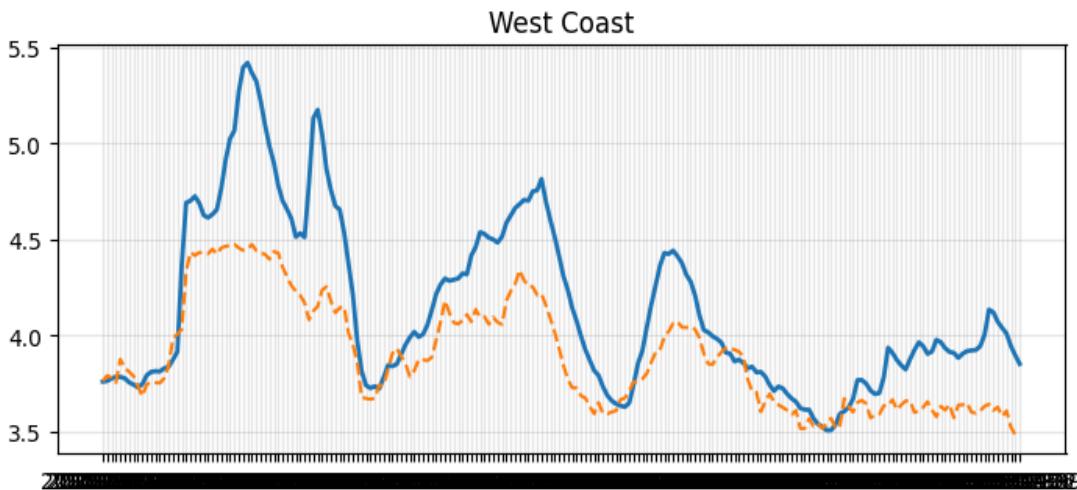


Figure 5.4 - Time series plot of weighted avg. predictions (orange dotted line) vs. Actual Price of Gas for West Coast

Despite the consistent underprediction, particularly during high-price periods, relative movements are captured well. Meaning we were able to consistently predict: when prices will rise or fall, and the slope of each movement. The fact that the most dramatic gaps occur during the periods of high volatility is very much consistent with the results and our conclusions from the West Coast SHAP analysis (see figure 5.1). We deduced that the region was the most sensitive to supply chain disruptions and internal supply outages. These events are not explicitly included in the features sources from the EIA. That is why the RF model struggled to fully capture the event-driven, dramatic shocks in this particular region.

Limitations

1. Regional data sparsity and noise

Many regional variables are reported inconsistently, leading to weak model performance.

2. Temporal mismatch

Some macroeconomic indicators (e.g., unemployment rate, CPI) originate from data that only include monthly reporting. Transforming monthly time series data to match weekly

values of the entire view reduced the clarity of the signal from these features.

3. Impact of events

A key limitation of the modeling approach is its inability to fully capture the influence of rare but high-impact events that have a drastic influence on gasoline prices. These events create structural breaks in the underlying economic relationships, causing models trained on historical patterns to perform poorly during or immediately after such disruptions.

Because these disruptions fall far outside the data distribution, the model cannot generalize to them without explicit features capturing the full scope of their impact.

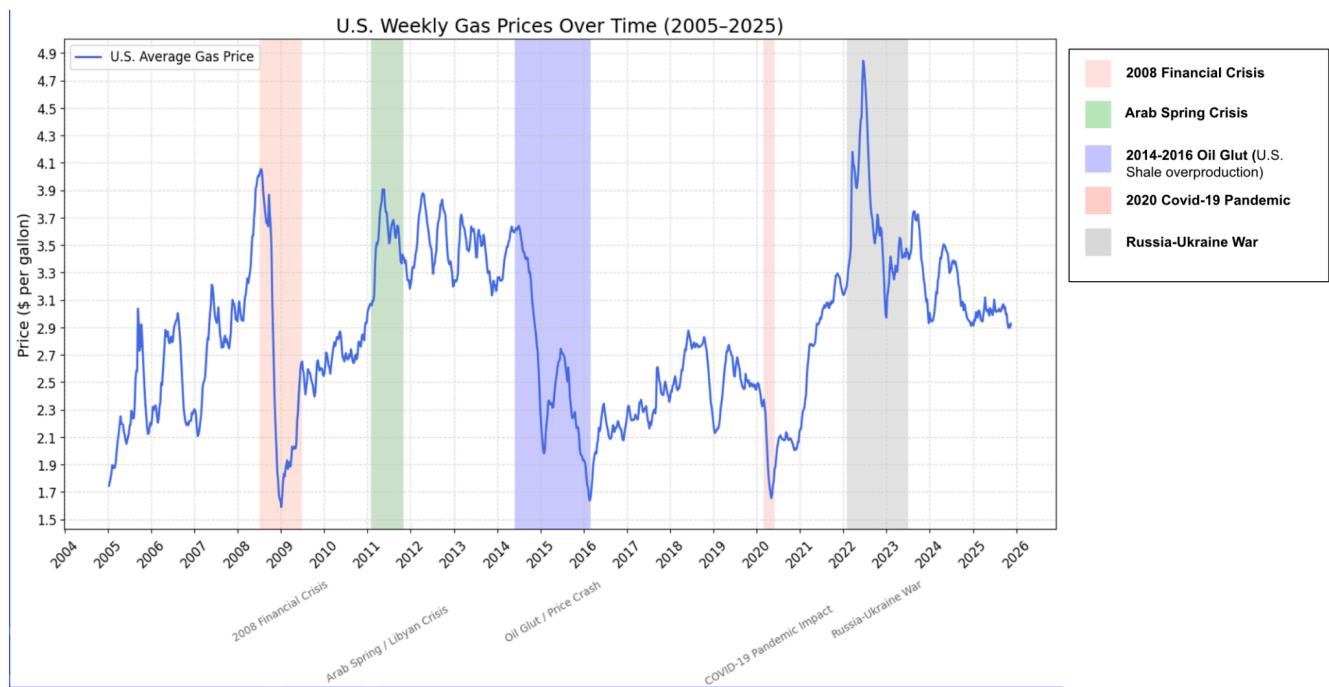


Figure 5.5- Time series plot of mean United States retail gasoline price annotated by high-impact geopolitical events. The plot illustrates the powerful impact these events have on prices. Prices are (relatively) stable under typical market conditions, but highly volatile during extreme disruptions, which are the periods when price forecasting is most valuable.

Global Demand Shocks (e.g., COVID-19 Pandemic)

The COVID-19 pandemic caused one of the largest and most abrupt collapses in gasoline demand in U.S. history. Weekly gasoline consumption fell over 40%, refinery utilization plummeted, and crude oil futures briefly traded below zero.

Models trained on “normal” economic conditions cannot anticipate:

- Nationwide lockdowns
- Sudden drop in transportation (people aren't driving)
- Supply chain disruptions
- Extremely rapid recovery swings

Geopolitical Events (e.g., Russia–Ukraine War, Middle East tensions)

Gasoline prices are highly sensitive to geopolitical developments affecting global oil supply. Examples include:

- Russian invasion of Ukraine
- OPEC decisions
- Sanctions/Tariffs on major oil exporters
- Conflicts in the Persian Gulf

These events change crude oil supply expectations rapidly and nonlinearly. Since they are not directly encoded in the datasets, the model responds only after their effects become visible in crude and product prices—introducing lag and reducing predictive accuracy.

Commodity Market Crashes (e.g., 2008 Financial Crisis, 2014 Oil Price Collapse)

Market crashes cause abrupt shifts in:

- Consumer demand
- Investor sentiment
- Refining margins
- Global transportation activity

These events introduce a breakdown of prior relationships between predictors and gasoline prices. Linear models such as Lasso are particularly vulnerable because they assume stable, additive relationships that do not hold during crashes.

Overproduction (e.g., 2011–2015 U.S. Shale Expansion)

U.S. shale production expanded dramatically during the 2010s, resulting in a period of domestic crude oil and natural gas oversupply. The result was a period of very low crude oil prices, and therefore retail gasoline prices. These events:

- Evolve over several years
- Are driven by investment behavior, not weekly indicators
- Permanently alter global supply dynamics

The model lacks features related to oil rig count, drilling investment, or future production outlooks. The model is unable to attribute price movements to longer-term supply changes.

6. Conclusion

This study demonstrates that forecasting gasoline prices benefits from multi-view modeling. Crude oil, macroeconomics, and regional factors each individually contribute to a segment of gasoline prices, but when combined, they paint a fuller picture. The final fused model through a weighted LASSO-based approach achieved an RMSE of 0.0629, MAPE of 0.0823, and an MAE of 0.2253, indicating that the model is able to approximate weekly U.S. retail gas prices with a small margin of error. These performance metrics reflect the value of a heterogeneous data model.

Out of the 3 models, LASSO Regression, XGBoost, and Random Forest, LASSO performed the best. LASSO's superior performance can be attributed to its ability to isolate the most important features, such as RBOB prices, CPI, and gasoline stock, while also reducing noise from less important features.

While LASSO gave us a satisfactory performance, the study also had key limitations. The United States is such a large country made up of many regions, which makes gathering the necessary data from all regions difficult. Another limitation is the inability to predict large, geopolitical events that have massive impacts on retail gasoline prices, such as financial crises or the COVID-19 pandemic. These events disrupt relationships between predictors and gas prices, causing sharp deviations from expected trends.

Despite these challenges, our findings point to several promising directions for future research. Increasing the breadth of regional supply data, such as weekly refinery reports or shipping constraints, could improve the predictive power of the regional view. Additionally, incorporating forward-looking market expectations may help the model anticipate sharp price swings, as opposed to reacting to them after the fact. Finally, a hybrid linear-nonlinear model may offer improved performance over a fundamentally linear model, due to the inevitability of nonlinear events.

Overall, this project illustrates that perfect prediction of weekly U.S. retail gas prices is impossible due to the volatile and multi-system nature of the gasoline market. However, substantial improvements to a predictive model can be achieved by fusing diverse information streams. Our multi-view modeling system provides a framework to combine global, domestic, and regional influences, delivering more accurate and stable forecasts than any single-model approach. As data availability and modeling techniques continue to evolve, multi-view fusion stands as an effective method for understanding and predicting complex market dynamics.

Use of Artificial Intelligence

As part of the project requirements, it was strongly recommended that we take advantage of generative AI tools to assist us in completing the project. However, we are required to document where artificial intelligence tools were used within the project. We want to make it clear that assistance from generative AI did not replace the analytical work or decision-making done by our team. Instead, it was referred to in order to enhance efficiency, provide technical guidance, and help validate certain choices.

Because our environment was hosted in Google Collabs Jupyter notebooks, we additionally used their built-in AI assistant, Gemini 3.0, for troubleshooting. Gemini was used specifically to debug errors and help correct syntax or logic issues that came up.

Our project was assisted primarily by ChatGPT 5.0, a popular generative AI chatbot. We used it to suggest relevant, quality data sources to integrate into our data views, suggest code for fusion and modeling purposes, and help produce visualizations to perform Exploratory Data Analysis and display our final results.

As challenges arose in the notebook environment, ChatGPT provided guidance on fixing merge and indexing issues when merging data by date/time index, as well as resolving mismatched time frequencies. It provided suggestions for handling NaN values or alignment problems. It clarified the correct use of certain modeling packages and tools like scikit-learn and SHAP. Referring to AI to handle these particular roadblocks helped the analysis to progress smoothly.

Although AI tools were leveraged extensively throughout the project, our ideas and interpretations of the project and its results were entirely our own. The AI tools served as a problem-solver and guide throughout the process, not as a substitute for our own work and personal ideas.