Biodiesel Commodity Price Forecasting Report (Year 2007-2024)

Executive Summary

This report investigates the structural drivers of weekly biodiesel price movements and develops a predictive framework to support short-term forecasting for commodity traders, biodiesel producers, risk management teams, and renewable energy policy analysts. Biodiesel prices are widely believed to be influenced by feedstock costs (particularly soybean oil), crude oil benchmarks (Brent), chemical inputs (methanol and ethanol), carbon pricing mechanisms (Carbon Tax and ETS), and broader macroeconomic indicators such as shipping activity (Baltic Dry Index). However, high volatility and strong interdependence across commodity markets make short-term price prediction challenging. The primary objective of this project is therefore to identify key economic drivers of biodiesel pricing and to build a forecasting model that meaningfully outperforms a naïve baseline assumption that next week's price equals this week's price.

Data were sourced from reputable institutions including the Food & Agriculture Organisation (FAO), the World Bank, Investing.com, and IA-USDA. The dataset was cleaned, standardized to weekly frequency, and enhanced with lagged variables (1–4 weeks) to capture autoregressive effects and short-term transmission mechanisms. Exploratory data analysis (EDA) conducted in Tableau revealed several important insights. First, global biofuel-related crop production exhibits long-term structural growth with cyclical fluctuations, particularly during periods of global stress such as the 2008–2009 financial crisis and the 2020 disruption. Second, an inverse relationship was observed between commodity production growth and biodiesel price changes, supporting the economic hypothesis that supply expansion exerts downward pressure on feedstock costs. Third, soybean oil prices demonstrated a stronger direct relationship with biodiesel prices compared to crude oil or methanol, reinforcing its importance as a primary cost driver.

Correlation and multicollinearity analysis revealed extremely high structural interdependence among agricultural and energy variables. Rather than removing correlated features, the modelling strategy prioritized regularized techniques capable of preserving economic realism

while stabilizing coefficients. Four modelling approaches were evaluated: Baseline (lag persistence), Ridge Regression, XGBoost, Elastic Net, and SARIMAX.

The naïve baseline model achieved an RMSE of 0.461 and R^2 of 0.837, confirming strong price persistence in biodiesel markets. Among all models tested, Ridge Regression delivered the strongest predictive performance, reducing RMSE to 0.402 (a 14.65% improvement over baseline) and increasing R^2 to 0.876. Elastic Net also improved upon the baseline (4.77% RMSE reduction), while SARIMAX provided modest gains (3.25% improvement) and confirmed strong autoregressive behavior. In contrast, XGBoost underperformed significantly, suggesting that biodiesel prices follow a predominantly linear and cost-driven structure rather than a highly nonlinear pattern.

Feature importance analysis from the Ridge model identified three dominant drivers: (1) previous week's biodiesel price, (2) cost of soybean oil, and (3) Brent crude oil price. These findings confirm that biodiesel pricing is largely persistence-driven but meaningfully influenced by feedstock and energy market fundamentals. Residual diagnostics further indicated that the Ridge model is well-calibrated, with errors centered around zero and no major structural bias.

Overall, the evidence supports the conclusion that biodiesel price formation is best characterized as a highly persistent, input-cost-driven linear process. Ridge regression is therefore selected as the final recommended model due to its robustness, stability under multicollinearity, and superior out-of-sample accuracy.

For stakeholders, close monitoring of soybean oil prices, crude oil benchmarks, and short-term price momentum is recommended to inform procurement timing, hedging strategies, and policy planning. While the model provides strong predictive performance, limitations remain, including the exclusion of granular policy shifts, real-time demand indicators, and sudden structural shocks. Future enhancements may include integrating high-frequency policy data, volatility indicators, and sentiment-based inputs. The next step is to operationalize the Ridge framework into an automated weekly forecasting pipeline with rolling retraining and stress-testing to ensure long-term robustness under evolving market conditions.

1. Problem, Goals and Audience

1.1 Audience/Persona

The primary audience of this project will be targeted at commodity trading desks, biodiesel producers, risk management teams, and renewable energy policy analysts. These stakeholders can refer to this report and findings to do short-term weekly price forecasts to optimize procurement timing, manage margins, hedge exposure, and plan operational decisions.

1.2 Problem Statement

Weekly biodiesel prices are often rumored to exhibit volatility driven largely by feedstock costs (soybean oil), energy markets (Brent crude), input chemicals (methanol, ethanol), carbon policy mechanisms (Carbon Tax and ETS), and **broader macroeconomic conditions** such as shipping activity (Baltic Dry Index).

The core problem is that it has always been difficult for the stakeholders in predicting the future biodiesel pricing using historical data due to the highly volatile commodity market. Creating challenges for stakeholders to make future business and policymaking decisions.

1.3 Goals

The primary objective of this project is to identify and quantify the key structural drivers influencing weekly biodiesel price movements. While governmental policy shifts and political developments were not explicitly modeled due to data limitations, this study focuses on observable, data-accessible market variables to better understand biodiesel price dynamics.

By leveraging historical commodity, cost, and macroeconomic indicators, the project aims to improve predictive accuracy beyond a naïve persistence benchmark (i.e., assuming next week's price equals this week's price) and develop a feasible forecasting framework suitable for real-world application.

1.4 Criteria for Success

Success in this project is defined across 2 dimensions: statistical performance and economic relevance

1.4.1 Statistical Performance:

- Model performance demonstrated consistent and meaningful improvement over the naïve persistence benchmark (lag-1 assumption) under true out-of-sample evaluation.
- Out-of-sample **RMSE** and **R-square** value is materially lower than the baseline model.

1.4.2 Economic Relevance:

The model must also demonstrate economic interpretability and structural coherence.

2. Data Sources and Data Dictionary

Data sources are extracted from reputable sources such as the **Food & Agriculture Organisation of the United Nations** , **The World Bank Group** and the Commodities Price data were extracted from **Investing.com**.

2.1 Core Data Dictionary

- Biodiesel_Price ($/gallon): IA-USDA weekly biodiesel price (target variable).
- Biodiesel_lag1: Previous week biodiesel price (autoregressive feature).
- Brent_Price: Weekly Brent crude oil price.
- Ethanol_Price: Weekly ethanol price.
- Methanol Price ($/metric ton): Weekly methanol input price.
- Cost of Soybean Oil ($/gallon): Weekly soybean oil feedstock cost.
- Carbon tax: Annual carbon taxation rate.
- ETS: Emissions Trading Scheme price.
- Price_DBI: Baltic Dry Index (shipping activity proxy).
- Cereals, primary / Maize / Palm oil / Soya beans / Sugar cane: (Annual agricultural production indicators

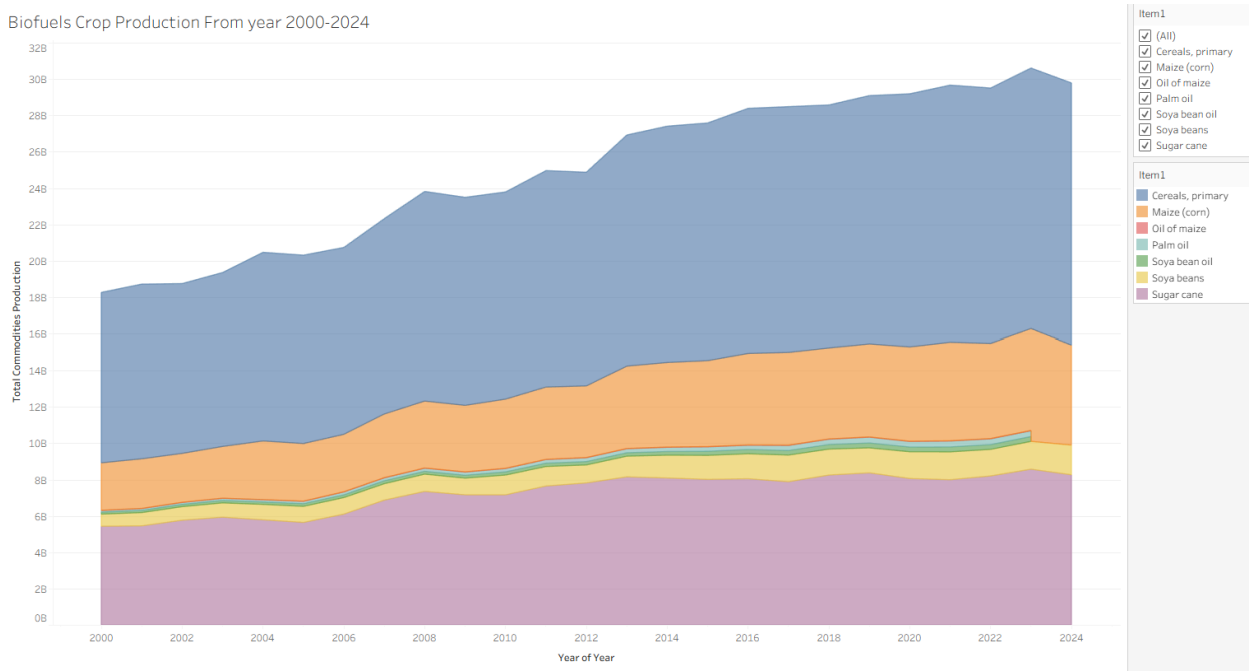
All datasets were cleaned to remove duplicate records, inconsistent date formats, missing observations, and redundant derived features. Lag variables of 1-4 were engineered to capture short-term autoregressive effects.

3. Patterns, Trends and Insights (EDA)

The EDA for this dataset was conducted on tableau to identify top level trends and patterns of the featured variables in the datasets for feature selection.

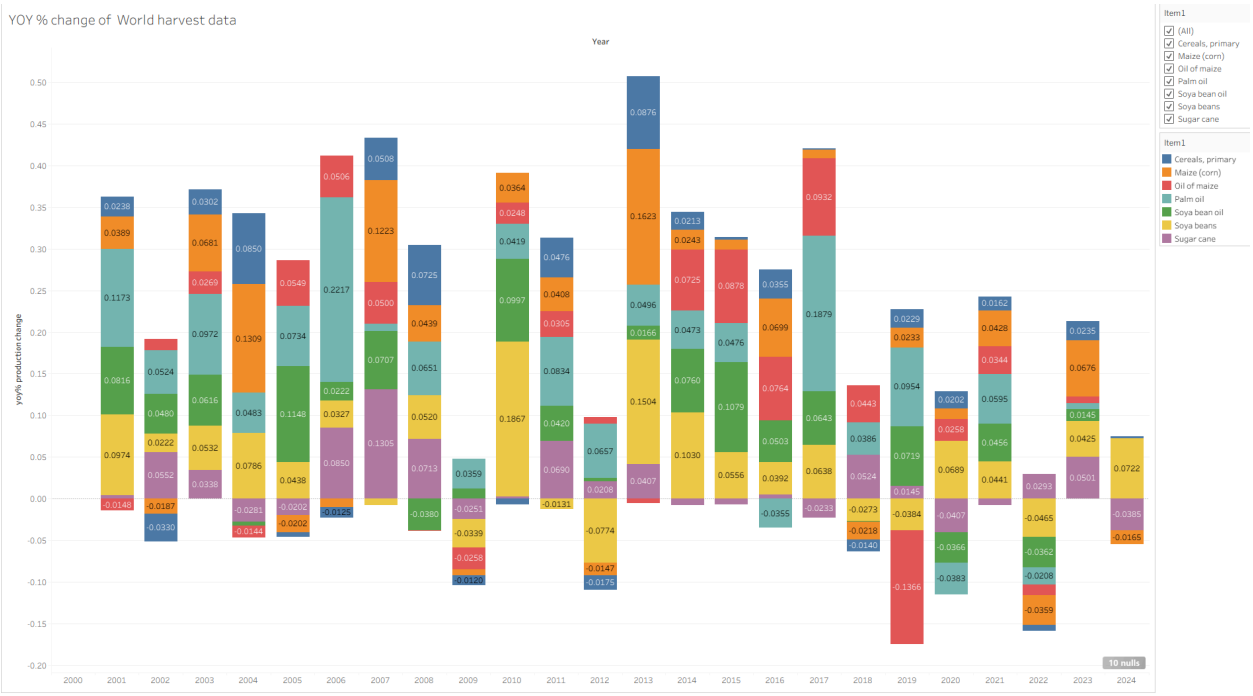
3.1 Supply Sided Trends Of Commodities

3.1.1 Biofuels Crops Product across years (2000-2024):



From 2000 to 2024, biofuel related crop production exhibits a clear long-term upward trajectory, this is a reflection of the structural expansion in global biofuel feedstock supply. Total production increases steadily over the period, with particularly strong growth observed between 2006 and 2014, suggesting a phase of accelerated agricultural scaling. Cereals (Grains) remain the dominant contributor throughout the sample period, accounting for the largest share of total production and driving much of the aggregate increase. Maize (corn) also demonstrates sustained growth, particularly after the mid-2000s, consistent with its central role in ethanol production. Sugar cane shows significant expansion during the early and mid-sample years before stabilizing at elevated levels, reinforcing its importance in global biofuel markets.
In contrast, other biofuels related crops such as palm oil, soya beans, and soya bean oil contribute to a smaller but steadily growing share, indicating diversification within feedstock inputs. With minor short-term fluctuations being visible that occur at the same time frame across all commodities such as a modest dip around the global financial crisis and slight softening toward 2024 the overall pattern remains structurally upward. This sustained expansion in feedstock supply has important implications for biodiesel price formation, as production volumes may influence input cost dynamics, lagged pricing effects, and broader commodity market interactions.

3.1.2 Biofuels Crops Product YOY % changes across years (2000-2024):

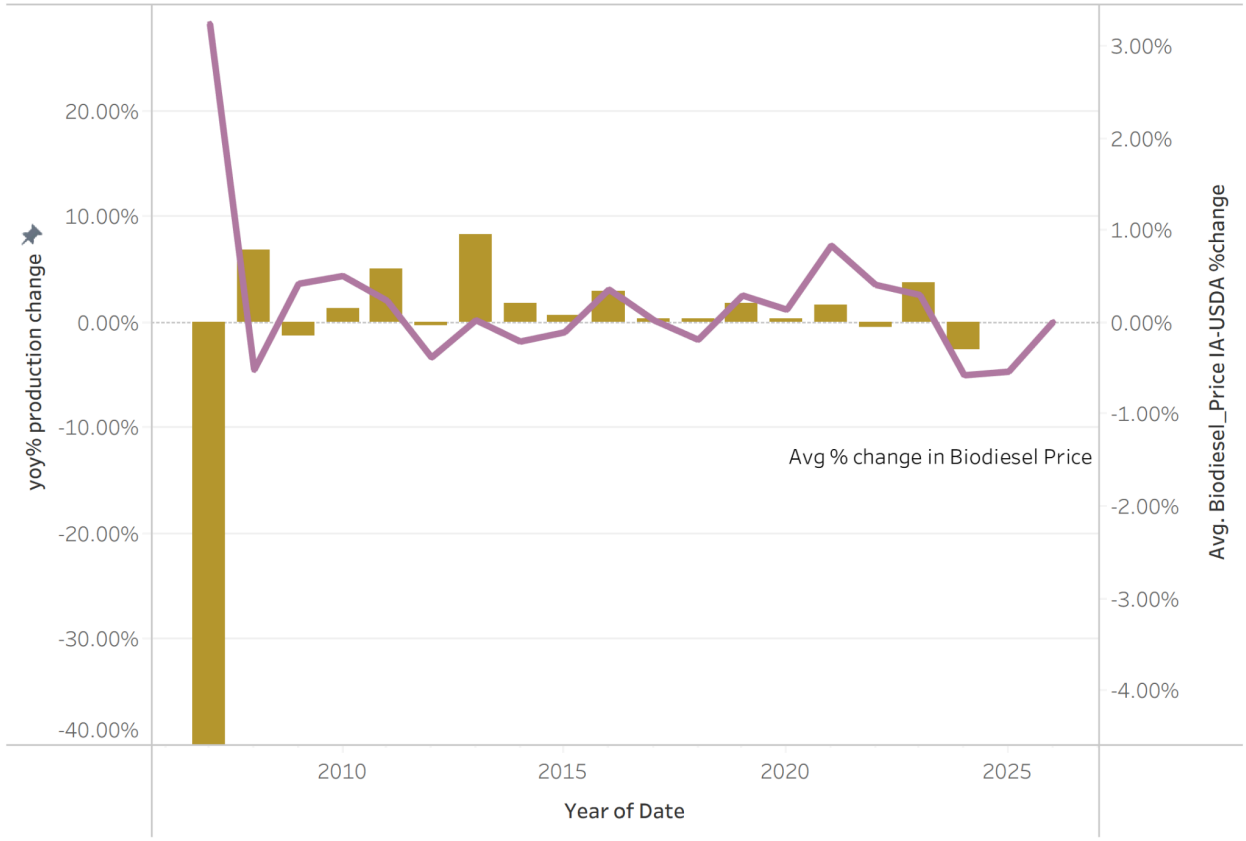


The year-on-year (YoY) percentage change in global harvest production reveals a pronounced **cyclical** pattern across major biofuel-related commodities, characterized by alternating periods of expansion and contraction rather than steady linear growth. While the long-term production trend remains upward, annual growth rates display significant volatility, particularly during periods associated with global economic or market disruptions. Visually significant slowdowns are observed more prominently appear around the 2008–2009 financial crisis, again in 2012, and more prominently during 2019–2020. This signals that crop production rates are also highly sensitive to macroeconomic stress, supply chain disruptions, and potential climatic variability. A broad-based rebound was observed in 2013, suggesting synchronized recovery dynamics across multiple feedstocks.

Among the commodities, oil-derived products exhibit relatively larger fluctuations, reflecting higher downstream processing sensitivity, while cereals demonstrate comparatively greater stability and serve as the structural foundation of global biofuel feedstock supply. These cyclical production dynamics have important implications for biodiesel price formation, as short-term supply shocks may exert lagged cost pressures on input markets.

3.2 Price Sided Trends Of Commodities with Biofuels Prices

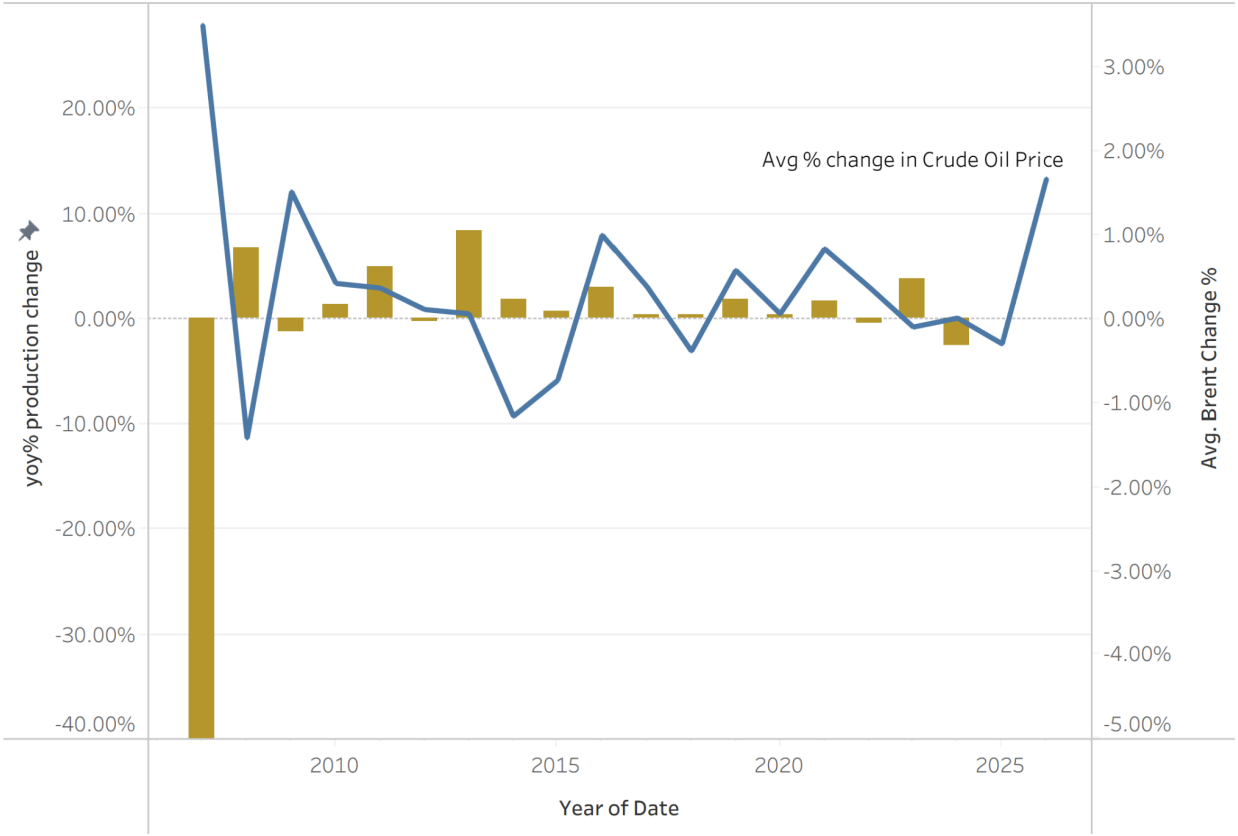
3.1.1 Commodity production VS Avg Yearly Biodiesel Prices % Change 2007 -2024:



The comparison between year on year (YoY) commodity production growth and average annual biodiesel price percentage change (2007–2024) indicates a generally inverse and lag-sensitive relationship between feedstock supply shocks and biodiesel price movements. Periods of strong production growth across major feedstocks such as cereals, maize, soya beans, and palm oil tend to coincide with moderating or negative biodiesel price growth. This suggests that supply expansion exerts downward pressure on input costs. Which makes plenty of economic sense.

Conversely, years characterized by weaker or negative production growth are more frequently associated with price increases, reflecting tightening feedstock availability. However, the relationship is not perfectly synchronous, implying that transmission effects may operate with temporal lags and may be influenced by additional cost components such as energy prices, freight dynamics, and processing margins. Overall, the graph supports the economic hypothesis that agricultural production growth functions as a structural supply-side driver within biodiesel price formation, reinforcing the importance of incorporating lagged production variables into predictive modeling frameworks.

3.1.2 Commodity Production vs Avg Yearly Crude Oil prices % change 2007 -2024:

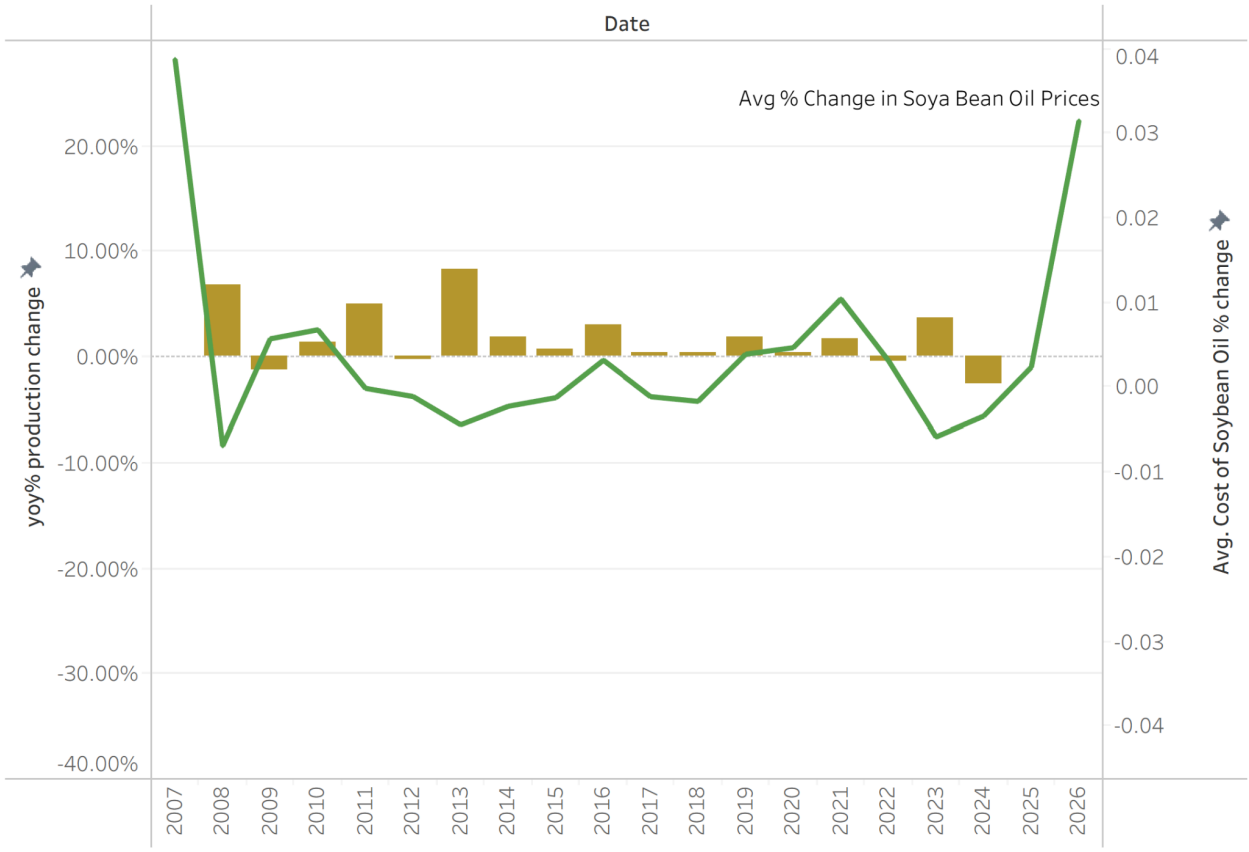


The comparison between year-on-year (YoY) commodity production growth and average annual crude oil (Brent) price percentage change from 2007 to 2024 suggests a relationship that is influenced more by macroeconomic energy cycles than by direct agricultural supply movements.

Crude oil price changes exhibit pronounced volatility, with sharp contractions during global stress periods (e.g., 2008 and 2014) and strong rebounds in recovery phases (e.g., 2009, 2016, and 2021).
In contrast, agricultural production growth remains relatively moderate and less synchronized with oil price fluctuations, displaying smaller amplitude changes and more gradual cyclical behavior.

The lack of consistent co-movement indicates that crude oil prices are driven primarily by global energy market conditions rather than immediate shifts in crop output. However, given biodiesel's linkage to conventional fuel markets, oil price volatility may indirectly influence feedstock demand and biofuel competitiveness rather than agricultural supply itself.

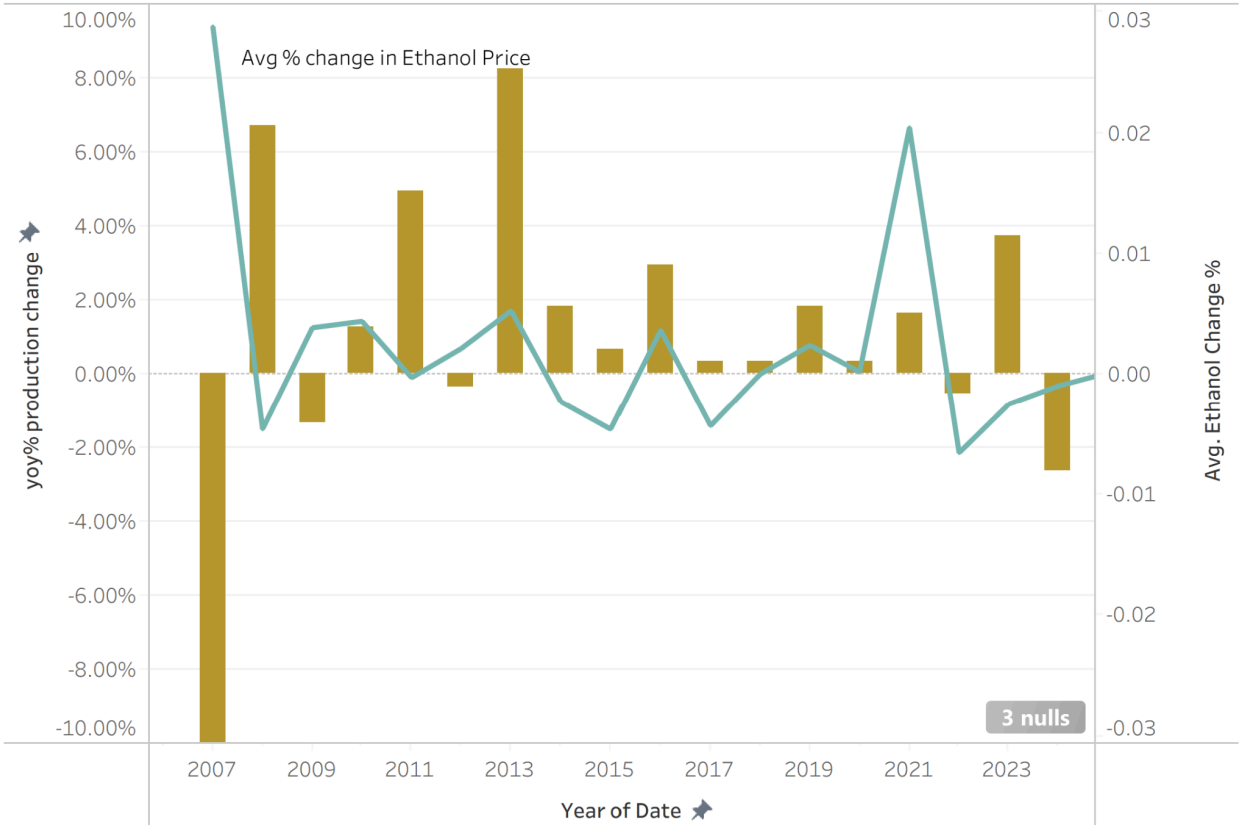
3.1.3 Commodity production vs avg yearly Soybean Oil prices % changes 2007 -2024:



The comparison between year-on-year (YoY) commodity production growth and average annual soybean oil price percentage changes from 2007 to 2024 suggests a more direct supply price interaction than observed with crude oil markets. Periods of stronger aggregate production growth particularly across soya beans, maize, and related oil products tend to coincide with moderating or negative soybean oil price growth, indicating the expected inverse relationship between feedstock supply expansion and price pressure.

Conversely, years marked by weaker or negative production growth are often associated with upward movements in soybean oil prices, reflecting tightening supply conditions. However, the relationship is not perfectly contemporaneous, as certain years display muted price responses despite production shifts, implying the presence of lag effects, global demand factors, and cross-commodity substitution dynamics. The pronounced volatility during disruption periods further underscores the sensitivity of soybean oil pricing to both agricultural output and broader market conditions.

3.1.4 Commodity production vs avg yearly Ethanol prices % changes 2007 -2024:

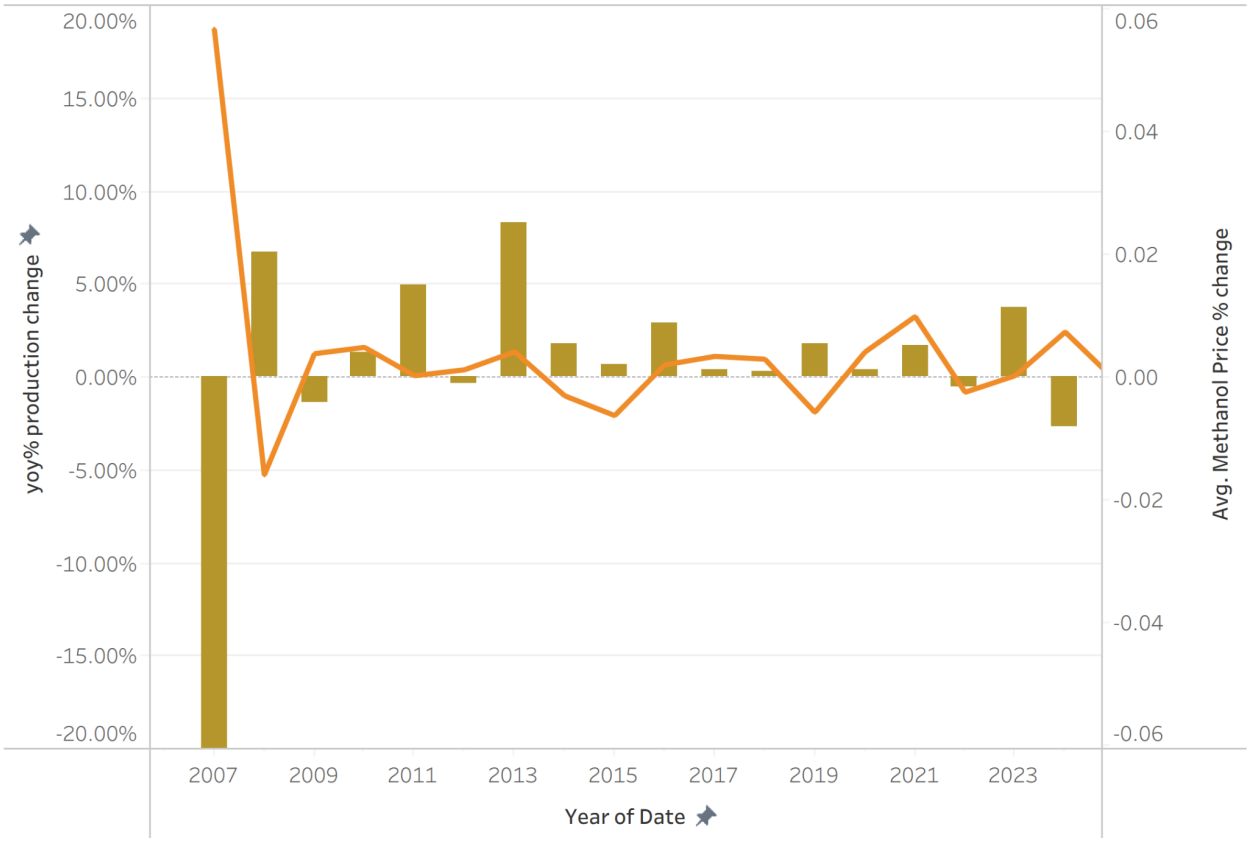


The comparison between year-on-year (YoY) commodity production growth and average annual ethanol price changes (2007–2024) suggests a generally inverse supply–price relationship, particularly for maize (red) and sugar cane (purple), the key ethanol feedstocks. Periods of higher production growth in these crops are typically associated with moderating or weaker ethanol price increases, indicating that greater feedstock availability places downward pressure on prices.

Conversely, slower or negative production growth often coincides with stronger price movements, reflecting tighter supply conditions.

However, ethanol prices also display independent volatility during broader market disruptions, suggesting that energy market dynamics and demand-side factors play an important role. Overall, the graph indicates that feedstock production is a relevant structural driver of ethanol pricing, though not the sole determinant.

3.1.5 Commodity production vs Avg yearly Methanol prices % changes 2007 -2024:



The comparison between year-on-year (YoY) commodity production growth and average annual methanol price changes (2007–2024) indicates a weaker and less consistent relationship than observed for biofuel feedstocks such as soybean oil or ethanol.

While certain periods show that stronger agricultural production growth coincides with softer methanol price movements, the pattern is not systematically inverse. Methanol prices exhibit distinct volatility that appears more aligned with broader energy and industrial market conditions than with crop supply dynamics. In particular, sharp price fluctuations during disruption years occur without proportional shifts in agricultural production, suggesting that methanol pricing is primarily driven by other fuel markets fundamentals rather than agricultural output.

3.1.6 Time Trends OF Fuel Prices 2007 -2024:

The time trend analysis of fuel price changes (Brent crude, methanol, ethanol, and biodiesel) indicates that while all fuel categories exhibit cyclical volatility, their long-term percentage change trends appear relatively flat to mildly declining over the sample period. Sharp movements are evident during major disruption years particularly around 2008 and again during the 2020–2021 recovery phase, highlighting sensitivity to global macroeconomic shocks and energy market instability.

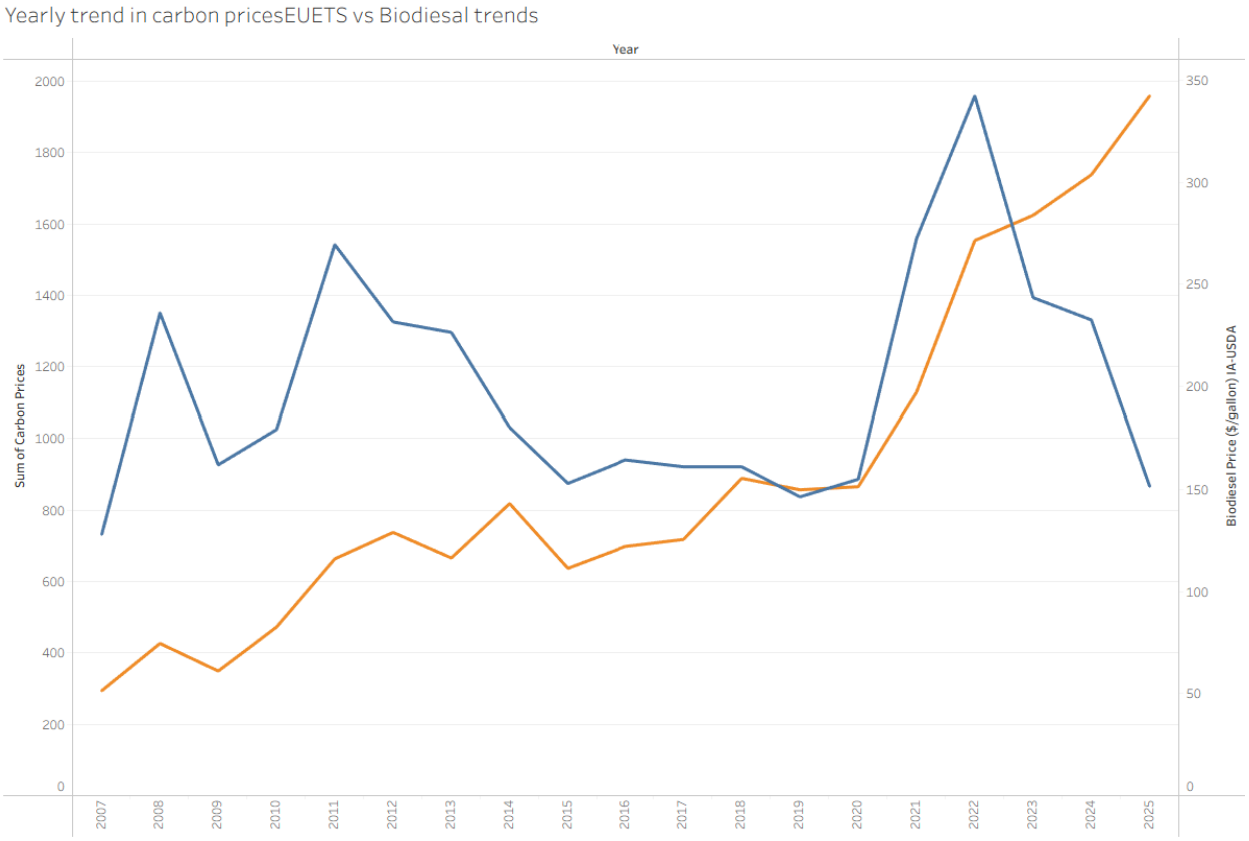
Brent crude and methanol display comparatively larger and more abrupt swings, reflecting exposure to global energy and industrial demand cycles.

Ethanol and biodiesel show similar but slightly more moderated movements, suggesting partial linkage to both agricultural feedstock dynamics and the broader energy complex.

The slight downward-sloping trendlines imply that extreme annual percentage changes have become less pronounced over time, pointing toward gradual market stabilization despite periodic shocks.

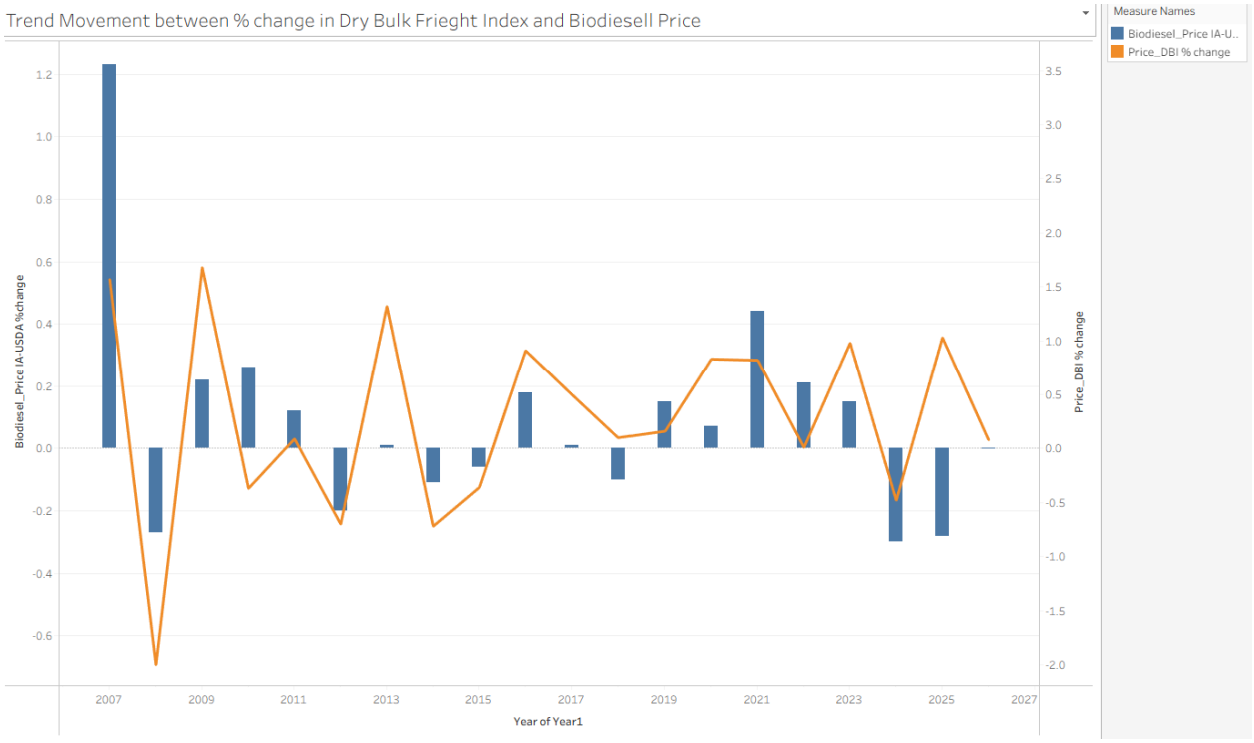
Overall, the graph supports the interpretation that fuel prices are strongly cyclical and externally driven, with biodiesel and ethanol influenced by both agricultural and conventional energy market forces.

3.1.7 Time Trends OF Fuel Prices 2007 -2024:



The comparison between EU ETS carbon prices and Crude Oil price trends shows a clearer positive co-movement in recent years, particularly from 2020 onward. While crude oil prices display cyclical fluctuations in the earlier period, carbon prices follow a more gradual upward trajectory before accelerating sharply after 2020. The simultaneous surge in both carbon prices and Crude Oil prices between 2020 and 2022 suggests that stronger carbon pricing mechanisms may enhance Biofuels relative competitiveness against fossil fuels as crude oil becomes more expensive.

3.1.8 Trend Movement between % change in Dry Bulk Freight Index and Biodiesel Price:



The comparison between percentage changes in the Dry Bulk Freight Index (DBI) and biodiesel prices indicates a partial but inconsistent relationship between freight cost volatility and biodiesel price movements. While certain periods such as recovery phases following major disruptions show directional alignment, the co-movement is neither stable nor consistently inverse.

Freight index changes exhibit sharp swings, reflecting global trade conditions and shipping market imbalances, whereas biodiesel price changes appear comparatively moderated and influenced by additional structural factors such as feedstock costs and energy prices.

Notably, large freight shocks do not always translate proportionally into biodiesel price changes, suggesting that transportation costs contribute to price dynamics but are not the dominant driver.

Overall, the graph implies that dry bulk freight costs function as a supplementary cost channel within biodiesel pricing, operating alongside broader commodity and energy market forces rather than determining price direction independently.

4. Predictive modelling and Results:

Following the preliminary exploratory analysis of the datasets, we observe that most supply-side variables across agricultural commodities and biofuel-related prices exhibit broadly similar directional trends over time, suggesting interconnected market dynamics. In contrast, EU carbon prices and the Dry Bulk Freight Index display more independent movement patterns, reflecting their linkage to policy and global trade conditions rather than direct agricultural supply cycles. Based on these observations, we proceed to conduct a formal correlation analysis to identify statistically significant relationships and select the most relevant variables for constructing a robust next-week biodiesel price prediction model.

4.1 Correlation of Variables

4.1.1 Multicollinearity Analysis:

A multicorrelation analysis was conducted on the existing variables and the below features pairing is found to be highly correlated.

Variable 1	Variable 2	Correlation
Soybean Oil_Price % change	Cost of Soybean Oil % change	1.0
Soybean Oil_Price (cents/lb)	Cost of Soybean Oil ($/gallon)	1.0
Methanol Price ($/metric ton)	Other Operating Costs ($/gallon)	1.0
Date	Year	1.0
Cereals, primary	Maize (corn)	0.99
Year	Palm oil	0.98
Date	Palm oil	0.98
Maize (corn)	Soya beans	0.98
Palm oil	Soya bean oil	0.98
Cereals, primary	Soya beans	0.97
Year	Cereals, primary	0.97
Date	Cereals, primary	0.97
Soya bean oil	Soya beans	0.97
Year	Soya bean oil	0.97
Date	Soya bean oil	0.96
Year	Maize (corn)	0.96
Date	Maize (corn)	0.96
Maize (corn)	Soya bean oil	0.96
Year	Soya beans	0.96
Date	Soya beans	0.96
Cereals, primary	Soya bean oil	0.95
Maize (corn)	Palm oil	0.95
Year	Carbon tax	0.94
Date	Carbon tax	0.94
Date	Sugar cane	0.85
Palm oil	Sugar cane	0.83
Soya beans	Sugar cane	0.8

The Correlation analysis revealed that several explanatory variables particularly agricultural production indicators exhibit high pairwise correlation. This is especially evident among crop-related variables such as cereals, maize, soybean derivatives, and related feedstock measures.

Where highly similar variables were identified such as the year-on-year percentage change in soybean harvest and the corresponding annual soybean harvest level. Only one representative variable was retained to avoid redundancy, given their strong conceptual and statistical overlap.

Accordingly, all percentage-change harvest variables were removed from the final dataset, and only the actual production harvest levels were retained, as these variables provided clearer structural trend information relevant to biofuel price dynamics.

However, this does not imply that all variables with high correlation represent identical data sources or redundant information.

In commodity markets, structural interdependence is expected. Agricultural outputs, feedstock prices, and energy inputs are economically linked through supply chains and substitution effects. Therefore, strong correlation among these variables reflects real-world economic structure rather than statistical error. From an economic perspective, such interdependence is logical and consistent with market

Given that the objective of this project is to model real-world biodiesel price formation, it would be inappropriate to remove variables solely on the basis of high correlation scores.

Artificially forcing independence among economically connected variables may instead:

- Distort the structural representation of the real world economic system
- Omit meaningful shared signal
- Introduce omitted variable bias
- Reduce predictive performance

It is also important to emphasize that correlation does not imply redundancy. Two highly correlated variables may still:

- Contain incremental predictive signal
- Improve model performance jointly
- Capture different transmission mechanisms within the commodity system

For example, agricultural production captures supply-side fundamentals, while input cost variables capture pricing transmission effects. Although related, these represent distinct economic channels.

Considerations for Model Selection:

However, the presence of multicollinearity does carry important methodological implications. Models that rely on pure coefficient interpretability such as Ordinary Least Squares (OLS) without regularization may produce unstable, inflated, or highly sensitive coefficient estimates in environments where predictors are strongly correlated. For this reason, alternative modeling approaches that are more robust to multicollinearity were prioritised. Regularized linear models such as Ridge regression were particularly appropriate, as L2 regularization stabilizes coefficient estimates while preserving the shared economic signal embedded in correlated variables. In addition, tree-based ensemble methods and SARIMAX specifications were explored as alternative predictive frameworks, given their structural flexibility and reduced sensitivity to linear dependency among features.

In summary, correlated variables were retained intentionally to preserve economic realism and predictive strength, while model selection was adapted to appropriately handle multicollinearity.

4.1.2 Variance Inflation Factor (VIF) Analysis:

After removing variables that were identified as redundant to our modelling objectives, a Variance Inflation Factor (VIF) analysis is further conducted to assess the degree of multicollinearity among explanatory variables. VIF measures how much the variance of a regression coefficient is inflated due to correlation with other predictors.

Variable	VIF
Cereals, primary	9,313.38
Sugar cane	4,652.40
Maize (corn)	3,842.53
Soya bean oil	2,855.73
Soya beans	2,249.99
Palm oil	1,067.58
Oil of maize	674.66
Carbon tax	365.26
Biodiesel_Price ($/gallon) IA-USDA	170.36
Cost of Soybean Oil ($/gallon)	138.17
Ethanol_Price	93.11
Brent_Price	90.07
Methanol Price ($/metric ton)	53.08
ETS	44.27
Price_DBI	6.04

Earlier, Ccrrelation matrix was used to screen for strong pairwise relationships and to confirm structural linkage across commodity variables.

In this component, the VIF was used to quantify multicollinearity severity in a regression context, showing how much predictors overlap when considered jointly.

The results revealed extremely high VIF values among agricultural production variables such as Cereals (9,313), Sugar Cane (4,652), and Maize (3,842). This reflects strong structural interdependence among crop-related variables, which is economically intuitive given their shared supply-chain and global production dynamics.

While high VIF values can destabilize coefficient estimates in Ordinary Least Squares (OLS) regression, they do not necessarily impair predictive performance. Rather, they signal that coefficients may not be individually interpretable in isolation.

Given the objective of this project is predictive modeling rather than causal inference, and given the use of regularized approaches such as Ridge regression, these correlated variables were retained to preserve system-level information.

4.2 Modelling Approaches

Time Series Data Preparation and Lag Structure:

Given that commodity prices exhibit strong temporal dynamics and underlying time trends, the dataset was reorganized to ensure appropriate time-based modeling. The data was first sorted chronologically and structured with a consistent weekly frequency to reflect the true temporal

sequence of market observations. This step ensures that model estimation respects the natural ordering of economic events and avoids information leakage across time.

To support next-week price forecasting, the target variable was defined as the one-step-ahead biodiesel price (t+1). This formulation ensures that predictions simulate real-world decision-making, where only current and historical information is available at the time of forecasting.

Lagged variables ranging from one to four weeks were introduced for both the target variable and selected explanatory drivers. These lag structures were designed to capture short-term transmission effects within commodity markets, reflecting the realistic delay between changes in input costs or market conditions and their impact on biodiesel pricing. Incorporating lag features allows the model to approximate the dynamic adjustment process inherent in energy and agricultural supply chains.

To ensure that there are no information leakages between the test and train model, the data set is further split based on chronological order of dates to also better extrapolate future price prediction.

By restructuring the dataset in this manner, the modeling framework accurately reflects the temporal nature of commodity markets and provides a fair and economically coherent basis for time series prediction.

4.2.1 Baseline Model:

A baseline model was established as the primary benchmark for evaluating all subsequent models and hyperparameter optimization efforts. This naïve benchmark assumes that the predicted biodiesel price for the following week is equal to the current week's observed biodiesel price.

Results by the baseline model:

- Baseline rmse:0.461
- R- square: 0.837

4.2.2 Ridge Regression Model:

Ridge regression was selected as a suitable modelling approach due to its ability to address multicollinearity through coefficient shrinkage. Unlike ordinary least squares (OLS), which can produce unstable and inflated estimates when predictors are highly correlated, Ridge regression introduces an L2 regularization penalty that constrains the magnitude of coefficients, thereby

stabilizing the model and improving out-of-sample performance. Rather than eliminating variables, Ridge retains all predictors while distributing explanatory weight more evenly across correlated features. This is particularly relevant for this dataset, where agricultural production variables, energy prices, and policy indicators exhibit extremely high correlation and elevated VIF values. By reducing variance without discarding structurally important variables, Ridge regression provides more reliable and interpretable estimates, making it well-suited for modelling biodiesel price dynamics in a highly interconnected commodity and energy market environment.

Results of the Ridge Regression Model:

	Baseline Model	Ridge Regression
RMSE	0.461	0.402
R-Square	0.837	0.876
% Improvement in RMSE Performance:	-	14.65%

From the above RMSE and R square comparison between Ridge modelling and the Baseline model, we can see that the Ridge model performs better by 14.65% in terms of its RMSE improvement which is considered a pretty good outcome within the commodity market. Furthermore, Ridge regression has a more ideal R-Square value which suggests that the selected variables explain about 87% of the price changes in biofuels. Where the model's variables improve the explanation of the biodiesel price trend 5% better than the baseline model.

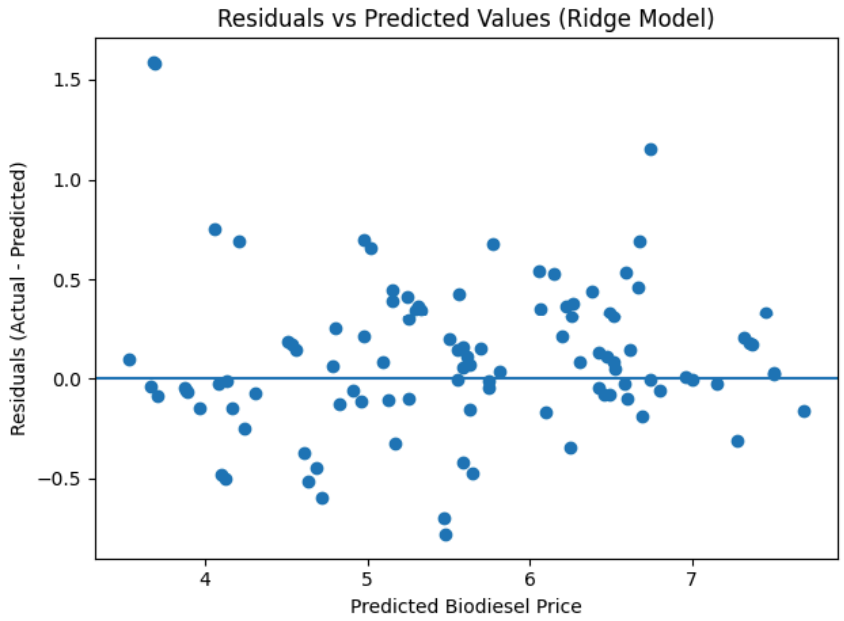
Feature	Coefficient
Biodiesel_lag1	0.75
Cost of Soybean Oil ($/gallon)	0.46
Brent_Price	0.22
Cereals, primary	0.06
Ethanol_Price	0.05
Carbon tax	0.03
Soya bean oil	0.02
Methanol Price ($/metric ton)_lag1	0.02
Methanol Price ($/metric ton)	-0.00

Oil of maize	-0.00
Sugar cane	-0.00
Price_DBI	-0.01
Methanol Price ($/metric ton)_lag4	-0.01
Ethanol_Price_lag4	-0.01
Maize (corn)	-0.02
Ethanol_Price_lag1	-0.02
Soya beans	-0.02
ETS	-0.03
Brent_Price_lag4	-0.04
Palm oil	-0.05
Cost of Soybean Oil ($/gallon)_lag4	-0.09
Brent_Price_lag1	-0.14
Cost of Soybean Oil ($/gallon)_lag1	-0.25

The model results indicate that the most influential predictors are Biodiesel_lag1, Cost of Soybean Oil ($/gallon), and Brent_Price, which are economically intuitive drivers of biodiesel pricing. Among these, Biodiesel_lag1 exhibits the strongest influence, highlighting the high degree of persistence present in weekly biodiesel prices. This finding is consistent with the strong R^2 observed in the naïve lag baseline model, confirming that current prices are heavily anchored to prior-week levels.

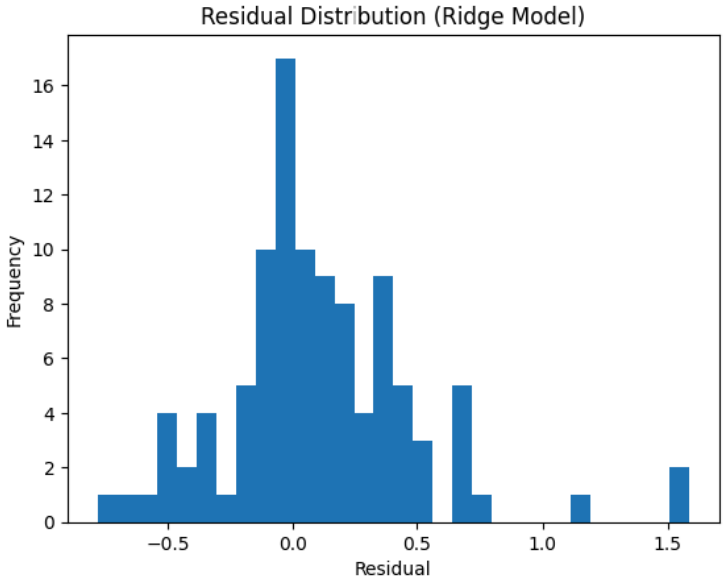
However, the superior performance of the Ridge regression model relative to the baseline demonstrates that biodiesel price formation cannot be explained by persistence alone. The inclusion of additional cost and energy market variables meaningfully improves predictive accuracy, indicating that input costs and broader crude oil dynamics contribute incremental explanatory power. In other words, while autoregressive momentum drives a substantial portion of price behavior, incorporating fundamental economic drivers enhances forecast precision and better reflects the structural mechanisms of the commodity system.

4.2.2.1 Ridge Model performance error analysis (Residual Scatterplot)



From the above residual scatter plots from the Ridge Model, residuals are approximately centered around zero with no strong systematic pattern across predicted values, indicating that the Ridge model captures the dominant linear relationships and does not exhibit strong bias. A small number of outliers remain, consistent with episodic market shocks that are difficult to explain using cost and macro drivers alone.

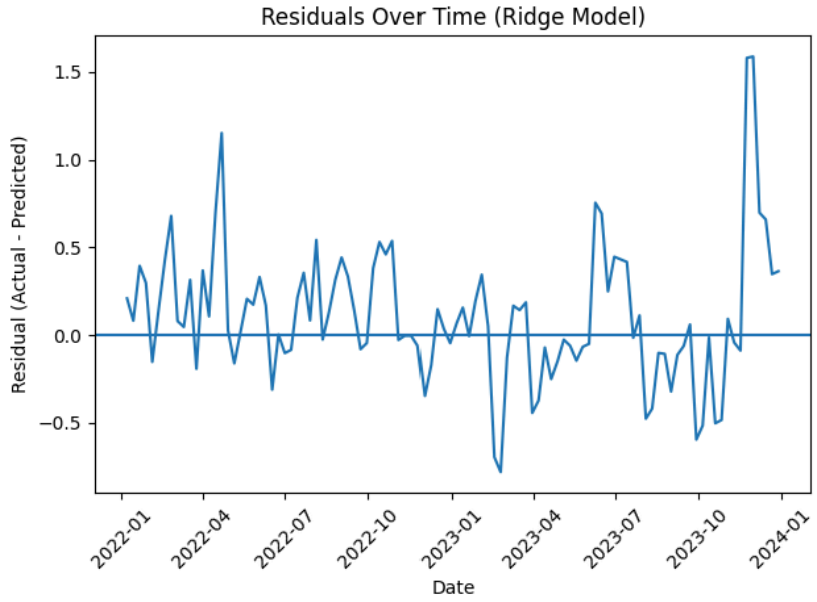
4.2.2.2 Ridge Model performance error analysis (Residual Distribution)



The residual distribution appears approximately centered around zero with a moderately symmetric shape, suggesting that the model does not exhibit systematic bias. While minor

right-skewness is observed due to occasional positive outliers, these likely reflect short-term market shocks rather than structural misspecification. Overall, the residual pattern supports the adequacy of the regularized linear modeling framework.

4.2.2.3 Ridge Model performance error analysis (Residual over time)



The residuals plotted over time appear largely centered around zero without a persistent upward or downward trend, indicating that the model adequately captures the underlying time persistence of biodiesel prices. While some clustering of residuals is observed during specific periods—particularly in recent years—these likely reflect market shocks or volatility regime changes rather than systematic model bias. Overall, no strong evidence of structural misspecification is detected.

4.2.3 XGBoost Model:

After implementing a regression model, XGBoost was introduced as a second modelling approach to evaluate how well the dataset performs under a decision tree-based framework. XGBoost (Extreme Gradient Boosting) is an ensemble algorithm that builds sequential decision trees, where each new tree corrects the errors of previous ones using gradient optimization.

Unlike linear models, it can capture nonlinear relationships, interaction effects, and threshold behavior without requiring explicit specification.

This is particularly useful in this dataset, where biodiesel prices may respond asymmetrically to commodity supply shifts, energy price volatility, and policy changes.

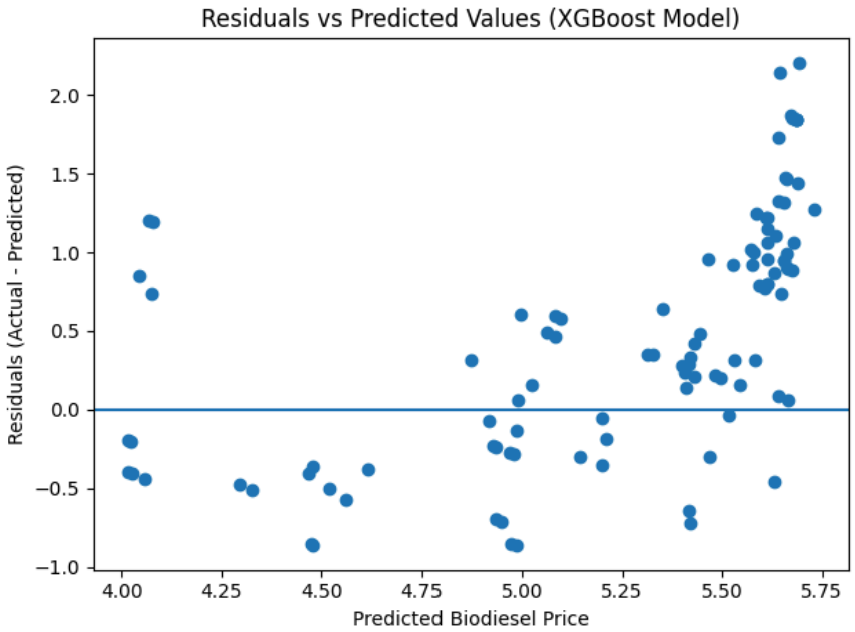
Additionally, XGBoost includes regularization mechanisms that reduce overfitting while maintaining strong predictive performance. Given the presence of multicollinearity and potentially complex dependencies among commodity and energy variables, XGBoost provides a robust comparative model to assess whether nonlinear structure improves next-week biodiesel price forecasting accuracy.

Results of the XGBoost Model:

	Baseline Model	Ridge Regression	**XGBoost Model**
RMSE	0.461	0.402	**0.884**
R-Square	0.837	0.876	**0.401**
% Improvement in RMSE Performance:	-	14.65%	**-91.75&**

Despite fine tuning the XG boost model, Ridge proves to be the better regression model in predicting the price of biodiesel 1 week later. The best model of XG boost is in fact worse than the naive model hence proving that the datapoint does not follow a non-linear trend and better follows a linear trend.

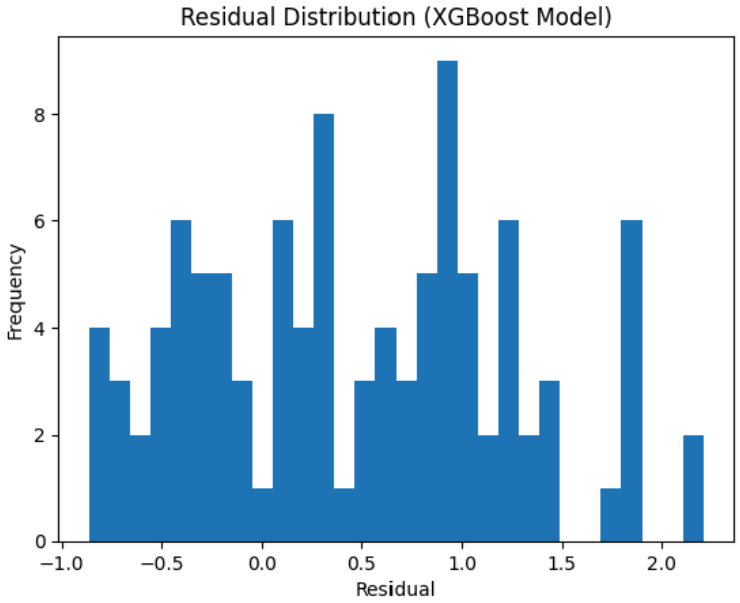
4.2.3.1 XGBoost Model performance error analysis (Residual Scatterplot)



The residuals vs predicted values plot for the XGBoost model reveals systematic underprediction at higher price levels and increasing error variance as predicted prices rise. This indicates heteroskedasticity and structural bias in the nonlinear model specification. In

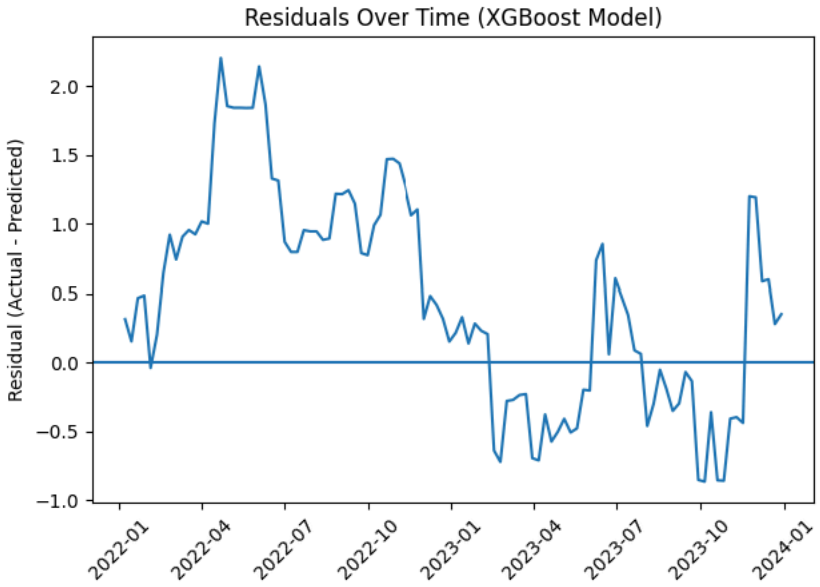
contrast to the Ridge model, the tree-based approach appears less capable of capturing the highly persistent and linear nature of biodiesel price formation.

4.2.3.2 XGBoost Model performance error analysis (Residual Distribution)



The residual distribution of the XGBoost model exhibits noticeable right skewness and heavier tails, indicating systematic underprediction during higher price regimes and greater sensitivity to volatility spikes. Compared to the Ridge model, the nonlinear tree-based specification demonstrates larger dispersion and structural bias, reinforcing the conclusion that biodiesel price dynamics are better characterized by a regularized linear regression.

4.2.3.3 XGBoost Model performance error analysis (Residual over time)



The residuals-over-time chart shows that the XGBoost model does not make random small mistakes. Instead, it makes similar types of errors for several weeks in a row. In 2022, the model mostly underestimated prices, while in early 2023 it tended to overestimate them. Toward late 2023, the size of the errors increased, showing that the model became less stable during more volatile market conditions. This suggests that XGBoost has difficulty adapting when the biodiesel market shifts, especially during periods of rapid change. As compared to the Ridge model, XGBoost is unable to adapt when market conditions become more volatile. Hence, validating the theory that Tree based Model is unsuitable for the Commodity Price prediction market.

4.2.4 Elastic Net Model:

The Elastic Net model is a linear regression technique that combines Lasso (L1) and Ridge (L2) regularization to improve predictive performance while addressing multicollinearity. By incorporating both penalties into the loss function, it performs variable selection through coefficient shrinkage while retaining groups of correlated predictors, making it more stable than Lasso alone when features are highly interrelated.

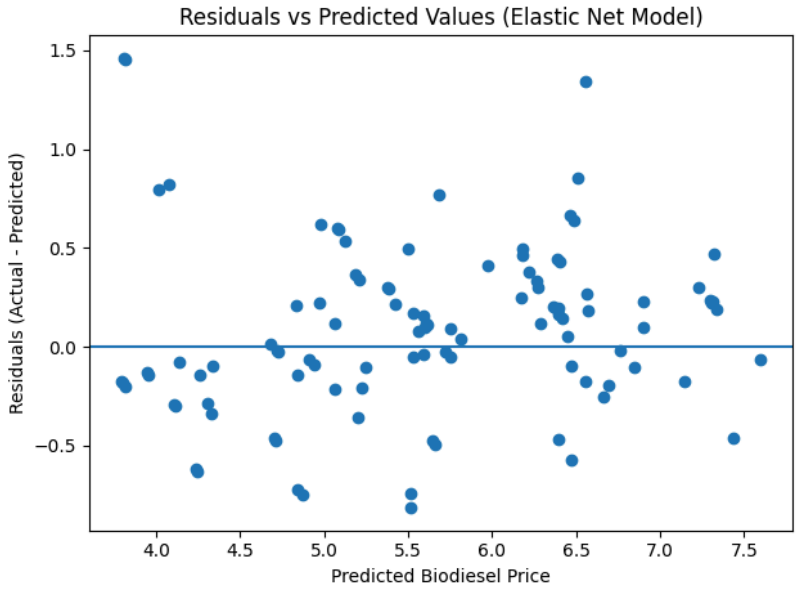
This is particularly relevant in this dataset, where agricultural production variables, energy prices, and policy indicators exhibit strong correlation structures. Given the interconnected and potentially nonlinear dynamics within commodity and energy markets, Elastic Net provides a structured, interpretable linear framework.

Results of the Elastic Net Model:

	Baseline Model	Ridge Regression	XGBoost Model	**Elastic Net Model**
RMSE	0.461	0.402	0.884	**0.439**
R-Square	0.837	0.876	0.401	**0.852**
% Improvement in RMSE Performance:	-	14.65%	-91.75%	**4.77%**

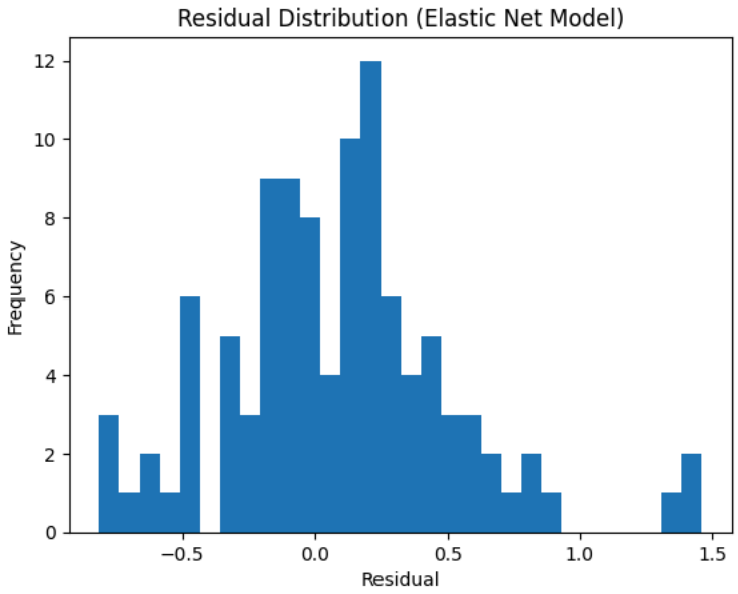
The Elastic Net model improves upon the baseline persistence model, reducing Test RMSE from 0.4612 to 0.439 and increasing Test R^2 from 0.837 to 0.852, representing a 4.84% reduction in prediction error and demonstrating that incorporating regularization and multiple explanatory variables enhances predictive accuracy beyond simple price persistence. However, the Ridge model performs more strongly, achieving approximately a 13.1% improvement over the baseline. This suggests that, given the high multicollinearity among commodity and energy variables in the dataset, uniform L2 shrinkage is more effective than partial feature selection, making Ridge the better-performing linear regularized model for next-week biodiesel price forecasting.

4.2.4.1 Elastic Net Model performance error analysis (Residual Scatterplot)



The Elastic Net model appears reasonably stable and well-calibrated. Errors are mostly random and centered around zero, with only moderate dispersion at extreme price levels. Compared to more volatile models (like XGBoost), this indicates stronger consistency and better handling of the persistent linear structure in biodiesel prices

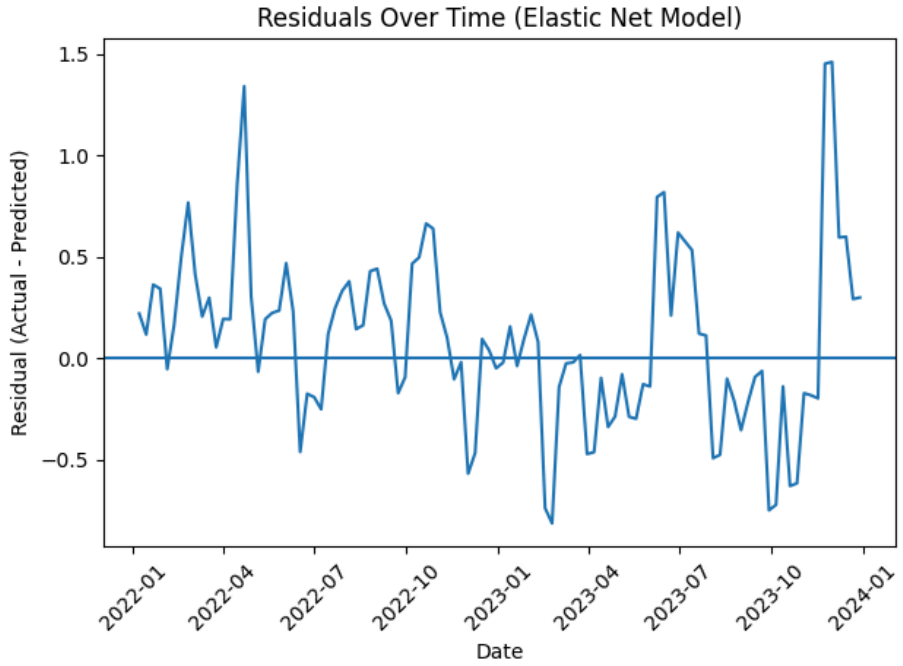
4.2.4.2 Elastic Net Model performance error analysis (Residual Distribution)



The Elastic Net model produces errors that are mostly small and centered around zero, indicating stable and reliable performance. While there are a few larger underprediction cases, the overall distribution suggests that the model captures the core structure of biodiesel price

movements reasonably well. However, in terms of the Rsquare statistics, the Ridge model still performs better in its prediction accuracy.

4.2.4.3 Elastic Net Model performance error analysis (Residual over time)



The Elastic Net model performs quite steadily over time. Most of its prediction errors are small and stay close to zero, which means it usually predicts prices fairly accurately. There are some periods where the errors move in the same direction for a few weeks, especially when the market changes quickly. However, the size of these mistakes does not become extreme. Compared to more complex nonlinear models, Elastic Net gives more consistent and reliable results, although it still finds it harder to predict during very volatile market periods.

4.2.5 SARIMAX Model:

SARIMAX is considered a suitable modelling candidate because biodiesel prices exhibit strong time-dependent behavior, including autocorrelation, seasonality, and lagged responses to external factors. SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) extends the traditional ARIMA framework by incorporating seasonal components and external explanatory variables into the time-series structure. It models the target variable using its own past values, past forecast errors, and differencing for stationarity, while simultaneously allowing external predictors—such as feedstock prices, energy prices, carbon taxes, and freight indices—to influence forecasts. This makes it particularly appropriate for this dataset, where biodiesel prices are shaped by both historical momentum and macroeconomic or commodity-based drivers.
Unlike purely machine learning models, SARIMAX provides interpretable parameters that explicitly capture temporal dynamics and lag structures, offering a structured econometric

framework that complements regression and tree-based approaches in forecasting next-week biodiesel prices.

Results of the SARIMAX Model:

	Baseline Model	Ridge Regression	XGBoost Model	**Elastic Net Model**	**SARIMAX**
RMSE	0.461	0.402	0.884	**0.439**	**0.446**
R-Square	0.837	0.876	0.401	**0.852**	**0.850**
% Improvement in RMSE Performance:	-	14.65%	-91.75%	**4.77%**	**3.25%**

Features importance is determined via the Coefficient assigned to each variable in the SARIMAX model.

Variable	**Coefficient**
Biodiesel_lag1	0.89
Cost of Soybean Oil ($/gallon)_lag1	0.09
sigma² (error variance)	0.03
Intercept	0.00
Brent_Price_lag1	0.00
ETS_lag1	-0.00
Carbon tax_lag1	0.00
Methanol Price ($/metric ton)_lag1	-0.00
Price_DBI_lag1	0.00

The SARIMAX results indicate that biodiesel price dynamics are primarily driven by strong autoregressive behavior, as reflected by the large coefficient on **Biodiesel_lag1 (0.89)**. This suggests a high degree of price persistence, meaning current biodiesel prices are heavily influenced by the previous week's price. Among the exogenous variables, only **Cost of Soybean Oil_lag1 (0.09)** shows a modest positive contribution, indicating limited but economically intuitive feedstock cost transmission. The near-zero coefficients for Brent price, carbon-related variables, methanol price, and freight index suggest that, within this specification and lag structure, their short-term impact is minimal relative to price momentum. Overall, the

model highlights that short-run biodiesel forecasting is largely persistence-driven, with external cost factors playing a secondary role.

Evaluation and Limitations of the SARIMAX Model:

While SARIMAX is theoretically well-suited for time-series modeling, it is sensitive to multicollinearity and lacks regularization. Given the severe collinearity among:

1. Soybean oil measures
2. Brent and ethanol prices
3. Carbon tax and ETS

Maximum likelihood estimation becomes unstable.

Ridge regression, by contrast:

1. Penalizes coefficient magnitude
2. Handles correlated predictors effectively
3. Directly optimizes predictive performance

Therefore, Ridge is statistically more robust in this high-dimensional, collinear commodity pricing environment.

4.3 Overall Model Evaluation Across: Baseline, Ridge, Tree Based Models, Elastic Net and SARIMAX

Across all models evaluated, the empirical results show consistency in indicatingg that weekly biodiesel prices are characterized by strong short-term persistence and a predominantly linear cost-driven structure. The naïve lag benchmark alone achieved relatively high explanatory power, confirming that current price levels contain substantial information about next week's price. However, among all tested specifications, Ridge regression delivered the strongest predictive performance, achieving the lowest RMSE and highest R^2 on the test set.

The superiority of Ridge regression reflects the structural properties of the dataset. Biodiesel prices are highly autoregressive and exhibit substantial multicollinearity across input cost variables, energy prices, and policy-related indicators. While the SARIMAX (ARX) framework successfully confirmed the underlying autoregressive structure and highlighted soybean oil cost as the most economically meaningful external driver, its lack of regularization made it more sensitive to multicollinearity and parameter instability. Similarly, tree-based ensemble models underperformed, suggesting that nonlinear interactions are not the dominant mechanism governing weekly biodiesel price movements.

Overall, the evidence supports the characterization of biodiesel price formation as a highly persistent, input-cost-driven linear process. In such an environment, regularized linear

regression, specifically Ridge regression offers the most robust and accurate forecasting framework by stabilizing coefficient estimates while preserving economically relevant signals.

Therefore, Ridge regression is selected as the final winning model for weekly biodiesel price prediction.

5. Model Recommendations

Based on the comparative modelling results, Ridge regression emerges as the most suitable predictive framework among the evaluated alternatives. It delivers the strongest out-of-sample improvement over the baseline model while effectively managing the high multicollinearity present across commodity, energy, and policy variables. Its L2 regularization stabilizes coefficient estimates without discarding structurally important predictors, resulting in a model that is both robust and economically interpretable.

Feature importance analysis further indicates that the three most influential drivers of biodiesel price fluctuations are:

(1) the previous week's biodiesel price (price persistence effect),

(2) the cost of soybean oil ($/gallon)

(3) Brent crude oil prices.

These findings suggest that biodiesel pricing is primarily shaped by short-term momentum and feedstock cost transmission, alongside broader energy market dynamics.

For traders, policymakers, and other market participants, close monitoring of these key variables is recommended to support pricing, hedging, and regulatory decisions. In particular, shifts in soybean oil input costs and crude oil benchmarks may serve as early indicators of biodiesel price adjustments.

6. Limitations

However, several limitations remained in this research. The model does not explicitly incorporate granular policy shifts, subsidy mechanisms, or real-time demand indicators, which may further enhance forecasting accuracy. Additionally, major unexpected changes in the market such as new government policies, sudden energy crises, or large global disruptions may change how biodiesel prices behave. When these big shifts happen, the relationships between variables may no longer follow past patterns, meaning the model may need to be updated or adjusted to reflect the new market conditions. Future work could include expanding high-frequency policy data, incorporating volatility indicators, or exploring hybrid modelling approaches that combine econometric time-series methods with machine learning techniques such as exploring LLM models in identifying market sentiment level from daily news.

7. Next Steps

As a next step, the Ridge regression framework should be operationalized into a production-ready forecasting pipeline with automated weekly data ingestion, preprocessing, and rolling retraining to ensure the model remains up to date with evolving market conditions. Additional robustness testing under alternative economic scenarios and stress environments should be conducted to evaluate performance stability during periods of volatility or structural shifts. Furthermore, trading firms and government agencies may consider extending their research into machine learning applications built around the Ridge framework, including adaptive or near–real-time updating mechanisms to better capture emerging trends and shifting market patterns.

Conclusion

This study set out to identify the key structural drivers of weekly biodiesel price movements and to develop a predictive framework capable of improving short-term forecasting accuracy beyond a naïve persistence benchmark. Through comprehensive exploratory data analysis, correlation assessment, and comparative modelling, the findings demonstrate that biodiesel prices are characterized by strong short-term persistence and a predominantly linear, cost-driven structure. Feedstock costs—particularly soybean oil—alongside Brent crude oil prices and prior-week biodiesel prices, were identified as the most economically meaningful contributors to price variation.

Among the models evaluated, Ridge regression emerged as the most robust and accurate forecasting approach. Its ability to manage severe multicollinearity while preserving economically relevant signals allowed it to outperform the baseline model and alternative machine learning and time-series frameworks. The results indicate that while biodiesel prices exhibit high autoregressive behavior, incorporating fundamental input cost and energy market variables provides meaningful incremental predictive power.

Overall, the empirical evidence supports the conclusion that biodiesel price formation is largely driven by persistence effects and input cost transmission within an interconnected commodity and energy system. A regularized linear modelling framework, particularly Ridge regression, is therefore well-suited to capturing these dynamics in a stable and economically interpretable manner. While limitations remain particularly regarding structural shocks and evolving policy environments, the developed forecasting framework provides a strong foundation for operational implementation and future refinement.

Done by: Joanna Woo

Date of completion: 27th February 2026