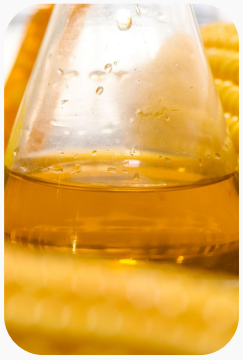
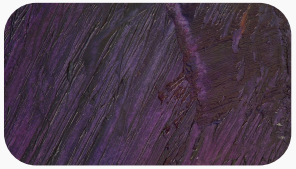
CAPSTONE
Final Presentation
- An Analysis of Biodiesel Price Fluctuations



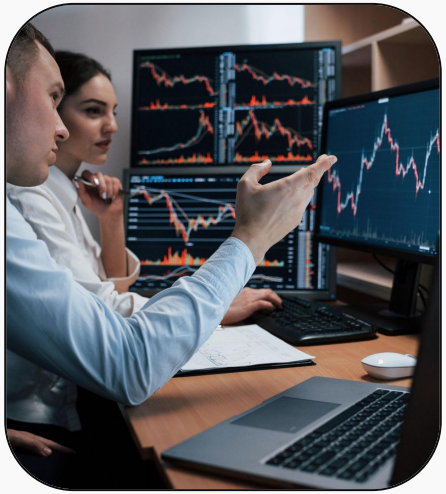



AGENDA

1. Persona
2. Goal
3. EDA
4. Data Sources
5. Modelling
6. Challenges & Limitations

Greetings To All..

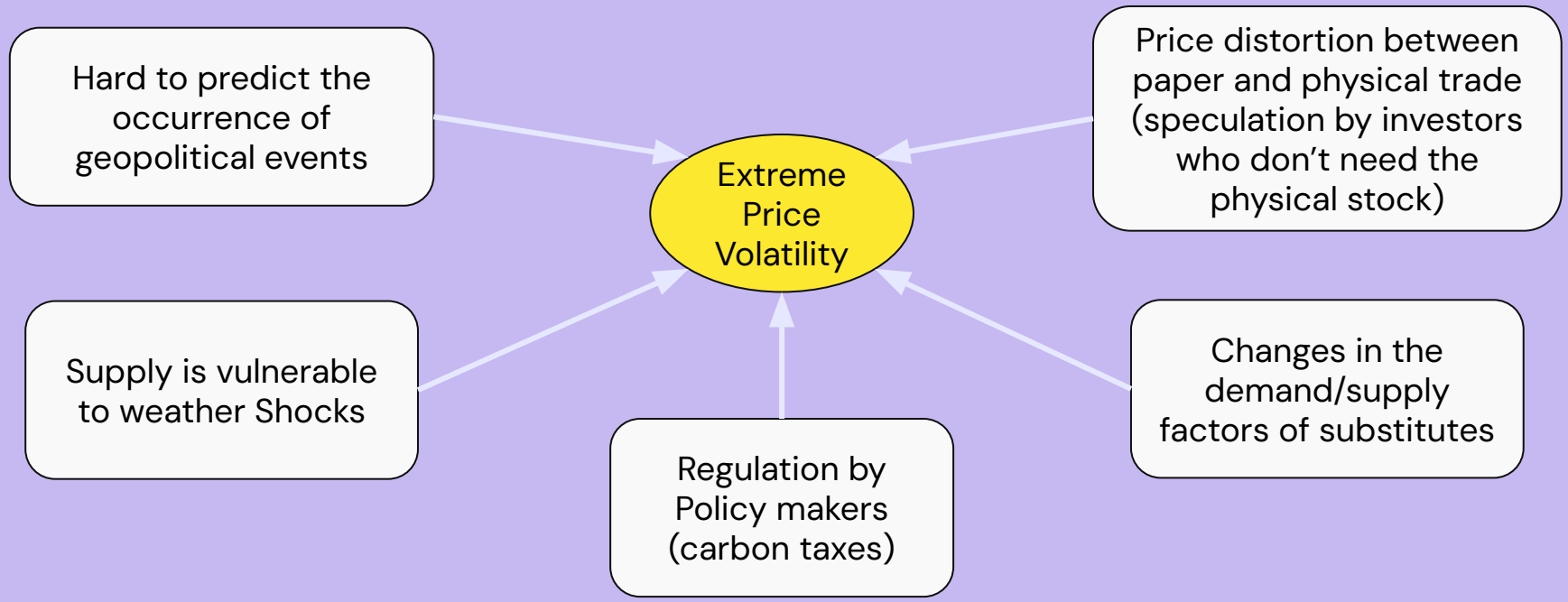
Economist & Policy Makers



Commodities Traders

Logistic and shipping operators

Pain Points of the commodity market....



Main Objectives:

“ To determine whether weekly biodiesel prices are driven predominantly by supply-side or demand-side factors ”

And...

“ To develop a predictive model to forecast short-term weekly price movements ”

Main Variables For Analysis:

Supply Factors

Harvest data of related commodities e.g. grains, corn, sugarcane,soybean and palm oil

Cost of Freight in Shipping Commodities e.g. Dry Bulk Index (DBI)

Demand Factors

Market prices of related commodities

Carbon tax prices across the years

Biofuels Crop Harvest Data

Ethanol

- Corn
- Sugarcane

Biodiesel

- Soybeans
- Palm Oil

Food

- Grain

Food & Agriculture Organisation of the United Nations

Biofuels & Crude Oil Price Data

Ethanol & Biodiesel

- Ethanol Futures Price: Investing.com
- Biodiesel Price:CARD or proxied estimate by 0.5(Soybean Oil) &0.5(Palm Oil Futures) : Investing.com

Crude Oil & Freight

- Brent Futures Price: Investing.com
- Dry Bulk Index (DBI): Investing.com

Carbon Eu/Ets

- State and Trends of Carbon Pricing: The World Bank Group (Dashboard here)

Let's Explore The Data Collected

A short intermission to Tableau!!

Highlights Of The Data Explored

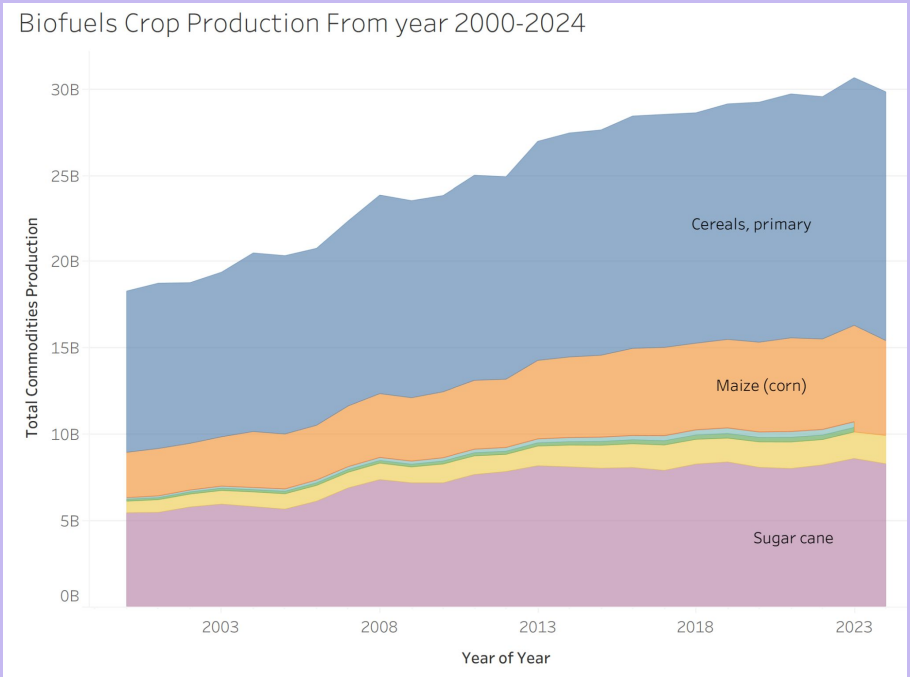


Fig 1: Yearly Harvest data on Commodities

Learning Point 1: The Top 3 most dominant crops being produced in the world are:

- Cereals(grains)
- Maize(corn)
- Sugarcane.

Top 2 producing countries for each dominant crops are:

Cereals(grains)

- China
- USA

Maize(corn)

- China
- USA

Sugar Cane

- Brazil
- India

Summary Of The Data Explored

Learning Point 2: China has emerged as the top Soya bean Oil producer overtaking America and Brazil from the year 2000-2023 & have retain their top position since 2010.

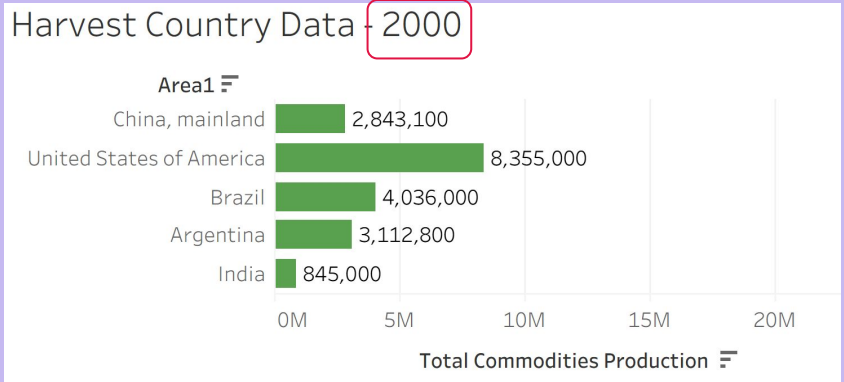


Fig 2: Harvest Country data on Soya bean Oil year 2000

VS

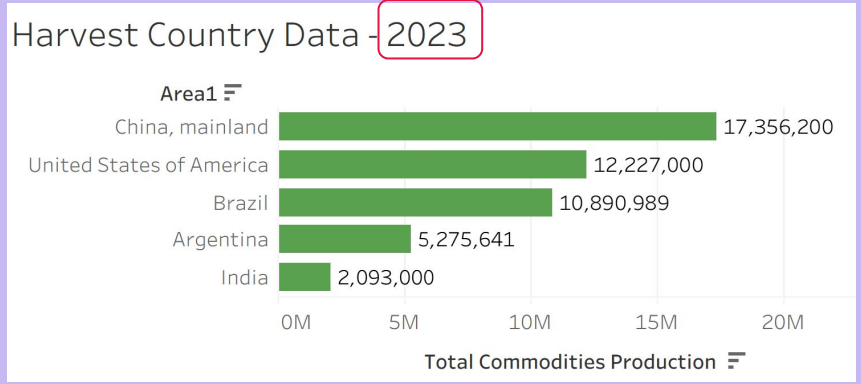


Fig 3: Harvest Country data on Soya bean Oil year 2023

Highlights Of The Data Explored

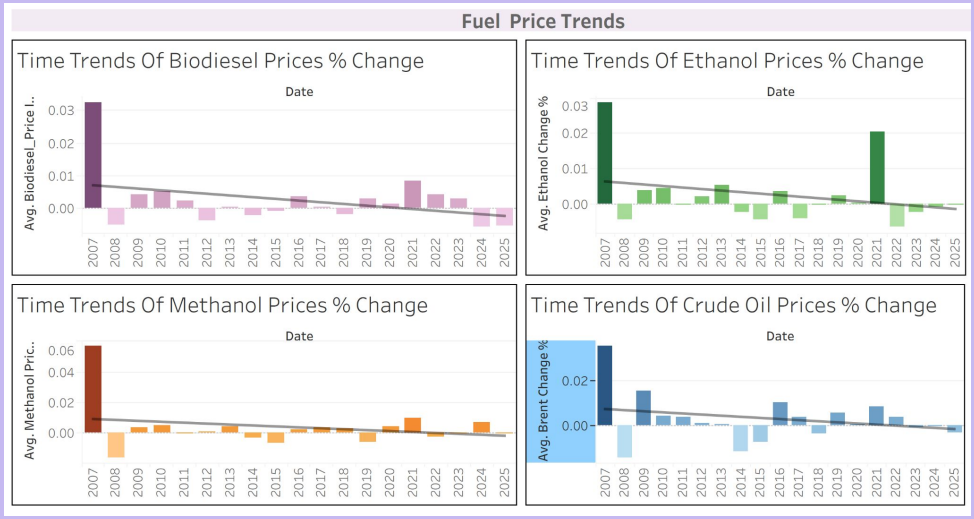


Fig 4: Fuel Price Trends

Learning Point 3: Fuel Prices and commodity production trends **moves largely in the same direction** and fuel prices seems to be more affected by cyclical macroeconomic trends (demand sided) then the supply side factors.

* This observation also serves as an early warning that the relationship should be examined more carefully for potential multicollinearity when developing the predictive model.

Summary Of The Data Explored

Learning Point 4: Global efforts to reduce carbon emissions have intensified, with the steepest carbon tax increases occurring after 2020. The full impact of carbon pricing may therefore extend beyond this dataset (2007–2024).

Fig 5: Carbon Tax Price data (year 2007)

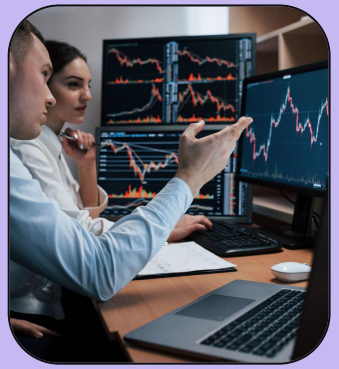
VS

Fig 6: Carbon Tax Price data (year 2024)

With this data....

Policymaker now can:

- Plan gradual tax escalation to avoid market shock
- Design carbon tax policy with awareness of inflationary spillover effects
- Anticipate impact on energy security and food supply



Commodities Trader now can:

- Adjust risk management strategies
- Anticipate impact on macroeconomic news in major producing countries
- Diversify exposure when production is geographically concentrated.

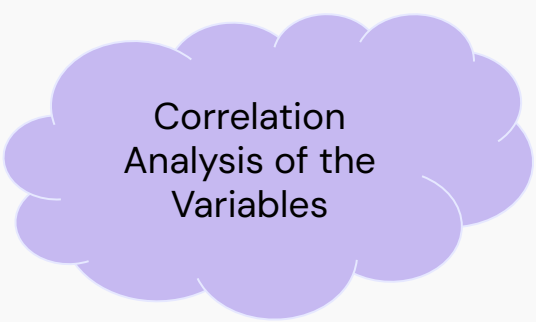
Logistic and Shipping Operator:

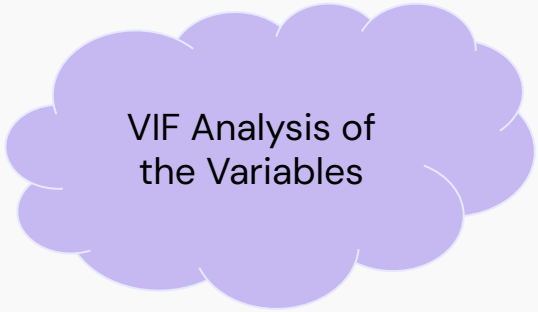
- Forecast freight demand based on crop production growth.
- Optimize trade routes based on dominant exporters.
- Mitigate risk of fuel costs against crude price spikes

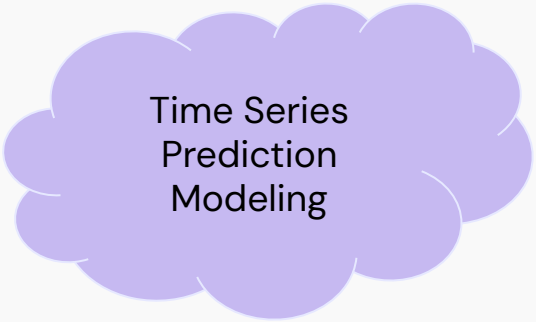
- Deeper variable analysis and multicollinearity assessment
- Identifying the ideal weekly prediction model.

Technical Analysis of The dataset

Technical Workflow:







4 models were tuned and tested to find the Winning Weekly Biodiesel Price Predictor model:

1. Ridge Regression Model
2. XGBoost Model
3. Elastic Net Model
4. SARIMAX Model

Correlation Analysis of Variables

Pairwise Correlation

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

Variance Inflation Factor (VIF)

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

- Strong correlations were observed within variable clusters (fuel-related indicators and crop harvest).
- Driven by common drivers within each group, the correlation is economically intuitive.
- The multicollinearity is economically justified, and as the variables capture related market dynamics, they are retained in the predictive model.
- Given collinearity, model selection must target methods that handle correlated features effectively.

Weekly Biodiesel Price Modeling Results

Ridge Regression

RMSE: 0.402
R-square: 0.876
% improvement from baseline: 14.65%

XGBoost

RMSE: 0.884
R-square: 0.401
% improvement from baseline: -91.75%

Elastic Net

RMSE: 0.439
R-square: 0.852
% improvement from baseline: 4.77%

SARIMAX

RMSE: 0.446
R-square: 0.850
% improvement from baseline: 3.25%

Baseline Model

RMSE: 0.461
R-square: 0.837

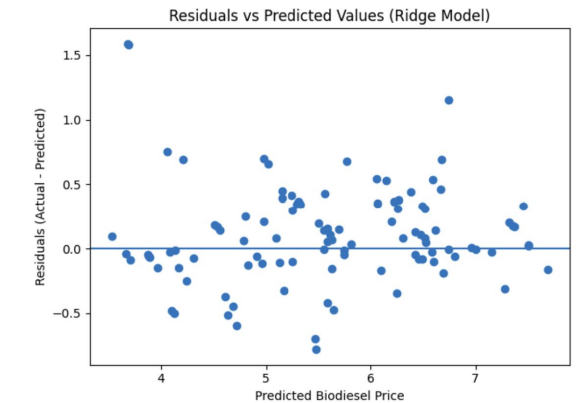
*Baseline Logic: Predicted biodiesel price for the following week = current week's observed biodiesel price.

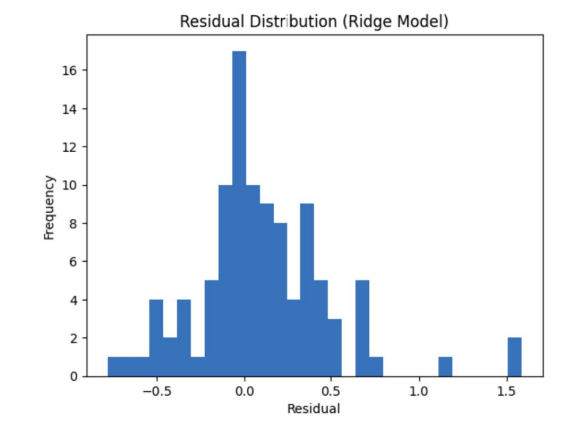
Ridge Model Evaluation

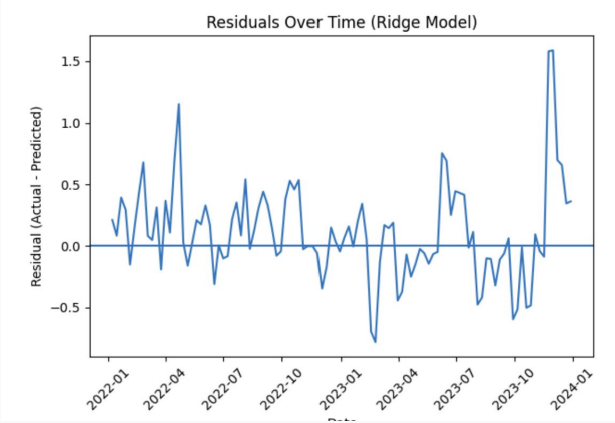
- Address multicollinearity through coefficient shrinkage by giving each variable a penalty.
- Model was finetune to include Lag variable to allow response time to occur over time

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

Ridge Model Evaluation







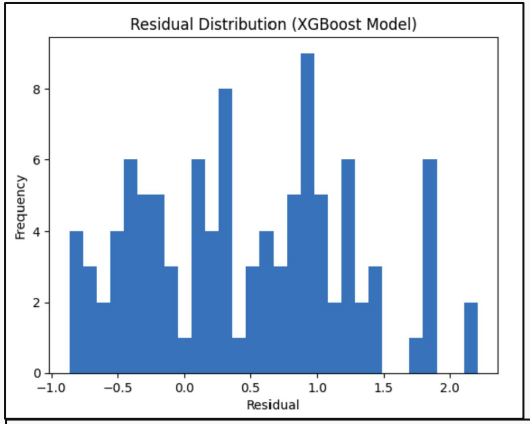
- Ridge model captures the dominant linear relationships and does not exhibit strong bias.

- Residual distribution appears approximately centered around zero with a moderately symmetric shape
- A little bit of right skewness is observed, like due to market shocks

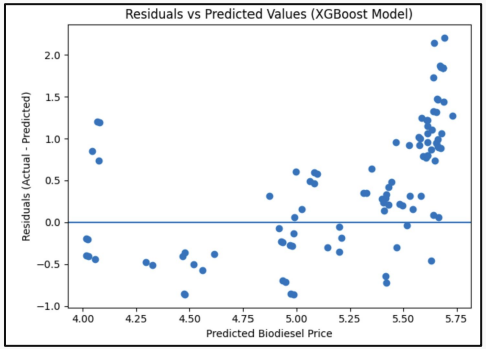
- Largely centered around zero without a persistent upward or downward trend
- Some clustering of residuals is observed during specific periods e.g due to market shocks

XGBoost Model Evaluation

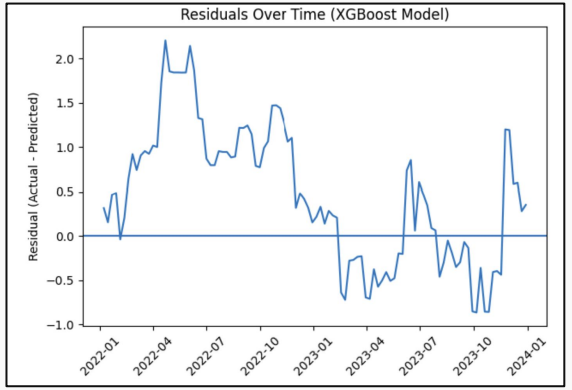
- Follow a decision tree-based framework.
- Includes regularization mechanisms that reduce overfitting



- The residual distribution exhibits noticeable right skewness and heavier tails, indicating systematic underprediction during higher price periods and greater sensitivity to volatility spikes.



- Residual over time, show the model making similar types of errors for several weeks in a row.
- Incapable of adapting as period shifts and market becomes more volatile.



- Residual over time, show the model making similar types of errors for several weeks in a row.
- Incapable of adapting as period shifts and market becomes more volatile.

The Commodity dataset is not suitable for a decision tree-base structure model.

SARIMAX & Elastic Net Model Evaluation

Elastic Net Model

- Another type of regularised linear regression model that takes in the feature of Lasso Regression and Ridge regression.
- As such, the residual results are similar to that of Ridge.
- However, results are slightly worse than the pure ridge model as features of lasso which the data was unsuited for was adopted.

SARIMAX Model

- A model that take season trends and moving averages into account
- Although results was slightly better then the baseline model,
- Many variables have to be removed especially the harvest dataset as harvest data was taken yearly why commodity prices are provided weekly.
- Sensitive to collinearity as such making this model overall less desired as compared to Ridge regression Model.

Weekly Biodiesel Price Modeling Results

Ridge Regression

RMSE: 0.402
R-square: 0.876
% improvement from baseline: 14.65%

XGBoost

RMSE: 0.884
R-square: 0.401
% improvement from baseline: -91.75%

Elastic Net

RMSE: 0.439
R-square: 0.852
% improvement from baseline: 4.77%

SARIMAX

RMSE: 0.446
R-square: 0.850
% improvement from baseline: 3.25%

Baseline Model

RMSE: 0.461
R-square: 0.837

*Baseline Logic: Predicted biodiesel price for the following week = current week's observed biodiesel price.

In Conclusion...

Top 3 Most Significant Variables For Biodiesel Price Volatility :

1. The previous week's biodiesel price (price persistence effect)
2. Cost of Soya Bean Oil
3. Brent Crude Oil Prices

Winning Predictive Model:

Ridge Regression 👑

Assumptions:

1. Ceteris paribus, biofuel prices are purely affected by only the identified variables and other unidentified factors are ignored
2. Model only analyses the trend from year 2002-2024 and assume that past relationships between variable remains stable continuously.

Next Steps...

The Most Influential Drivers of Biodiesel Price :

1. The previous week's biodiesel price (price persistence effect)
2. Cost of Soya Bean Oil
3. Brent Crude Oil Prices

Model Improvement Plan:

1. Further Add on to model's adaptability by connecting dataset to API address to get real time price information for prediction
2. Explore the usage of LLM model in Judging Sentimental index from news article to better capture the Macroeconomic Confidence Level.

Next Steps For Stakeholders...

Economist & Policy Makers:

- Implement carbon pricing gradually to control inflation risk
- Monitor energy transition impact

Commodities Traders:

- Use weekly forecasts to optimize hedging and inventory timing
- Manage correlated cluster exposure

Logistic and shipping operators:

- Hedge bunker fuel exposure
- Align fleet capacity with crop production cycles

THANK YOU!

Done by: Joanna Woo