



DBA5102

NUS-Cargill Analytics Innovation Challenge

Price Risk Modelling

Optimising Buying Behaviour for Edible Oils and FX Futures

Team Name: Finding Beta

Rachel Sng (A0231921N) | Susan Koruthu (A0231905L)
Widya Gani Salim (A0231857Y) | Wong Cheng An (A0232039M)
Felipe Chapa Chamorro (A0179033E) | Gino Martelli Tiu (A0231956Y)

The Business Problem

“ How does Cargill leverage machine learning to better understand customers and hedge risk?”



(1) CUSTOMER SEGMENTATION

Can we segment customer groups based on their attributes?

- ☐ Historical pricing data
- ☐ Transaction volumes
- ☐ Transaction frequency

Clustering Problem



(2) CUSTOMER TRANSACTION BEHAVIOUR

Can we leverage on internal and external data and use these to predict whether the customer...

- ☐ Transacts or does not transact
- ☐ The quantity of said transaction

Classification Problem



(3) PREDICTING & UNDERSTANDING OPEN MARKET PRICE

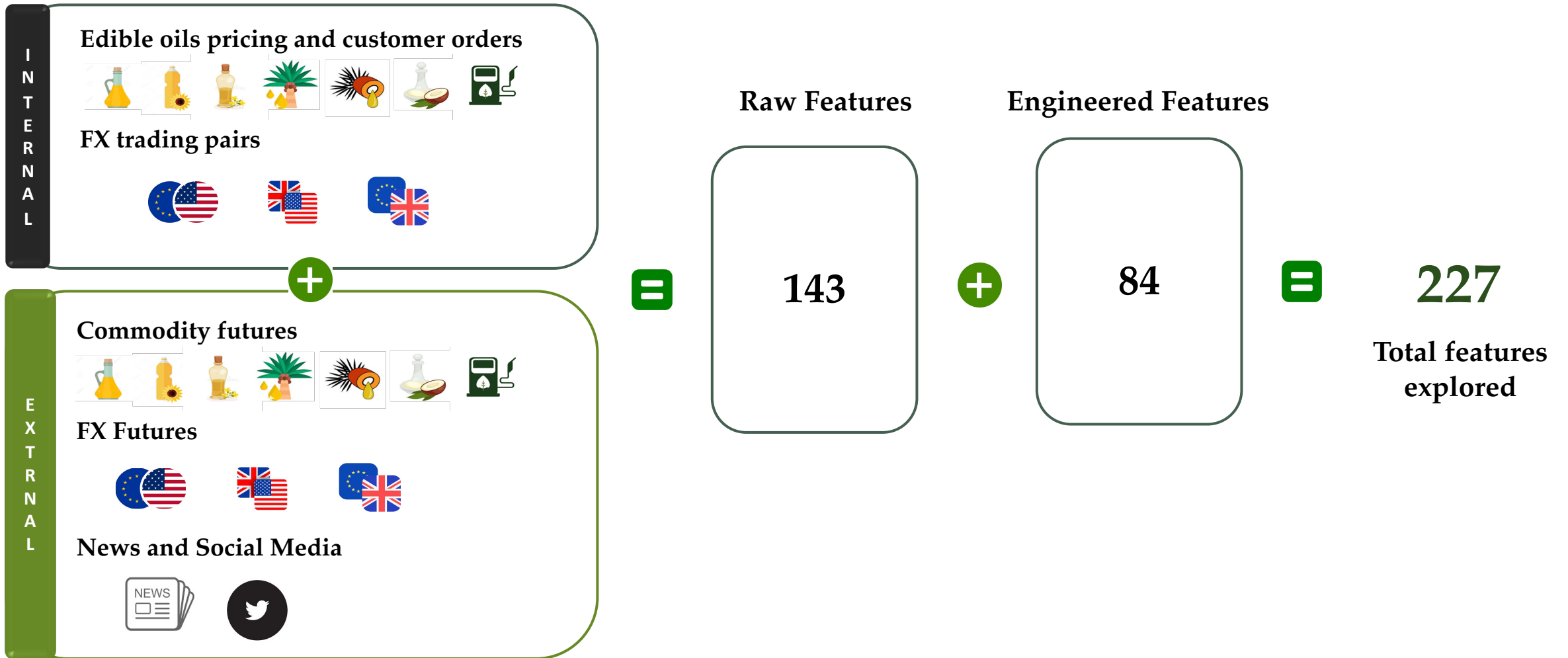
Knowing the customer buy, no buy decision and the foreseen quantity, can we:

- ☐ The open market price when the transaction occurs

Regression Problem

Data Sourcing and Enrichment

The below data sources were mined for variables that could be used in the three use cases outlined:



Multiple data sources used in feature engineering

In particular, the following were derived on top of data sourced -



Time Variables:

- Month
- Quarter



Sales Metrics:

Source: Cargill Customer Transactions

- Sales quantity
- Cumulative sales
- Sales frequency
- Month since sale
- % sales on month



Commodity Price & Contract Features

(FX, edible oils, crude)

Source: BarChart getHistory API

- Closing price
- Closing price % growth
- Open interest
- Open interest % growth
- Volume
- Volume % growth



News sentiment :

Source: BarChart getNews API, Twitter, GoogleNews API

- Crude
- Oils
- Ethanol
- Grains



Customer Transaction Behavior:

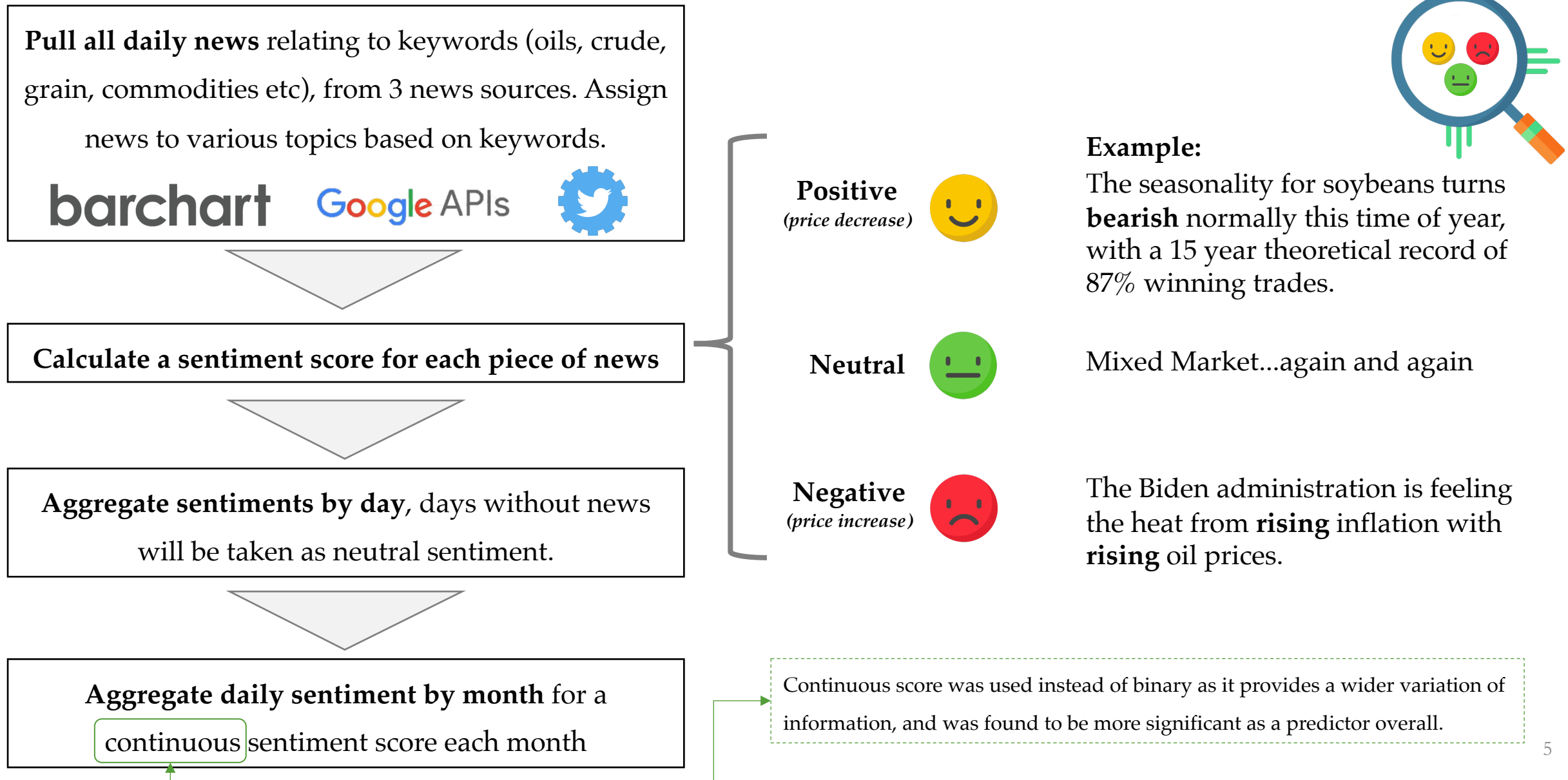
Source: Cargill Customer Transactions

- Oil volume per coverage period
- Contracts transacted per period
- FX exposure



Full Data Dictionary in Appendix 1

Sentiment Analysis: Wide variety of external sentiment aggregated to a monthly estimate of overall market sentiment towards prices





Part I: Customer Segmentation

Identifying clusters of customers with similar purchase behavior

Clustering Motivation

Original Customer Transaction Dataset

(454,336 transaction records, 28 feature columns)

1

Remove noisy data

- No pricing quantity
- Partially cancelled and erroneous transactions



229,405 records

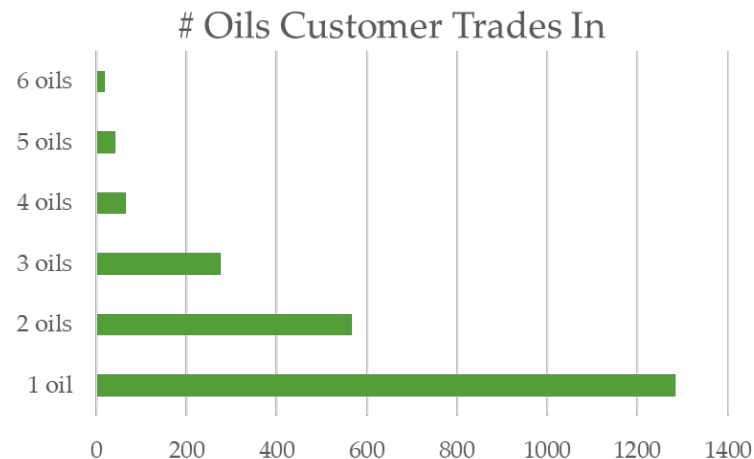
2

Aggregate sell side data to customer level

- 98% of total cleaned transactions



2,110 unique, 6 oils



How should we think about clustering?



CUSTOMER BEHAVIOR

1. Sales Contracts
2. Oil Quantities Sold
3. Currency Risk



over...



TIME HORIZON

1. Short Term: 0-5+ Months
2. Medium Term: 6-12 Months
3. Long Term: 12+ Months

What are the potential applications?



- Cold start problem & initial customer profiling
- Timing of customer transactions based on profile
- Profile based sizing of positions & coverage period
- Understanding of FX exposure – customer behavior relationship, if any.

Clustering Approach

For each oil...

1

Clustering methods:

Tried based on data shape and outlier detection

- K-Means [*BEST METHOD OVERALL*]
- DBSCAN
- OPTICS
- Hierarchical clustering (dendrogram)

2

Remove mean and scale to unit variance:

Reduce impact of extreme scale features before distance or variance-based methods (PCA, clustering)

3

Dimensionality reduction – PCA:

Reduce impact of dimensionality on distance-based clustering methods

- 90% Information retained
- Reduced to ~10 dimensions

4

Evaluation:

Determine number of clusters to maximize intra-cluster similarity and minimize inter cluster similarities

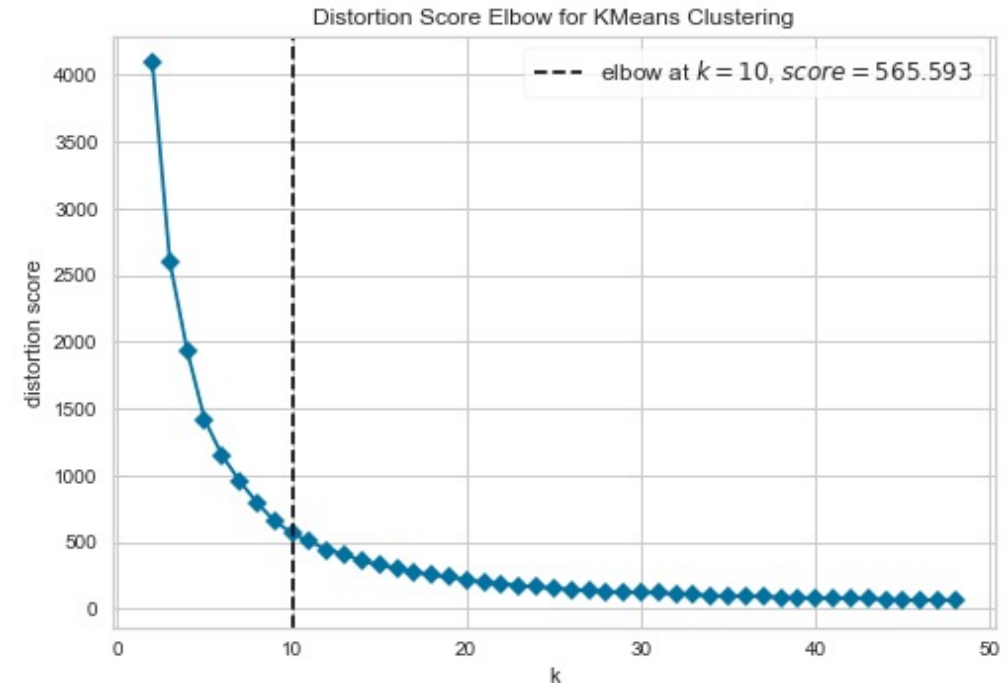
- K-Elbow method
- Silhouette score

5

Iterative feature selection:

Iteratively try different combinations of internal and external features to evaluate if there are meaningful patterns separating clusters

- e.g. Most insightful features were mostly internal and FX pairs were ultimately found to not be useful



Palm Oil | Silhouette score: 0.68

Clustering Results: Clear Groups of Short, Mid and Long Term Sales Behavior

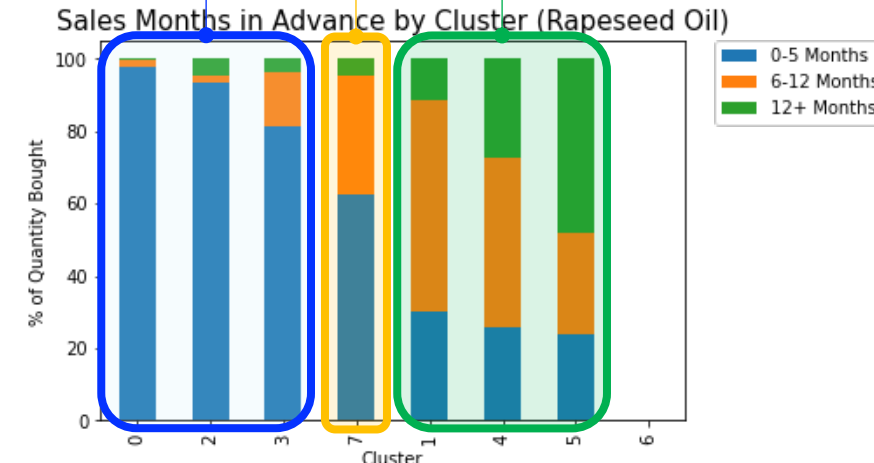
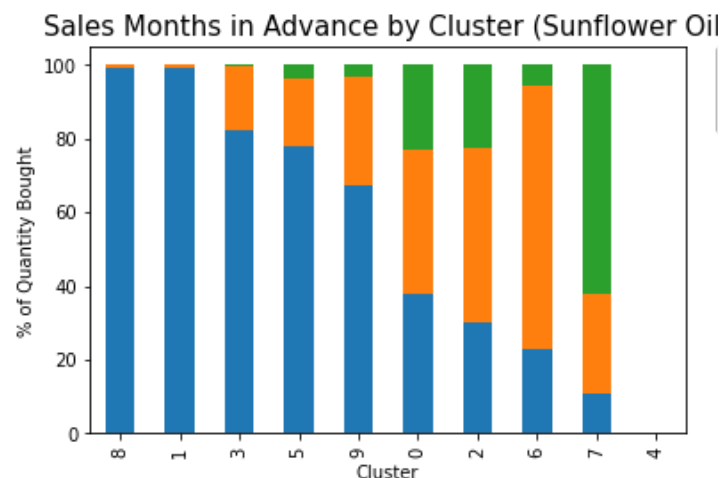
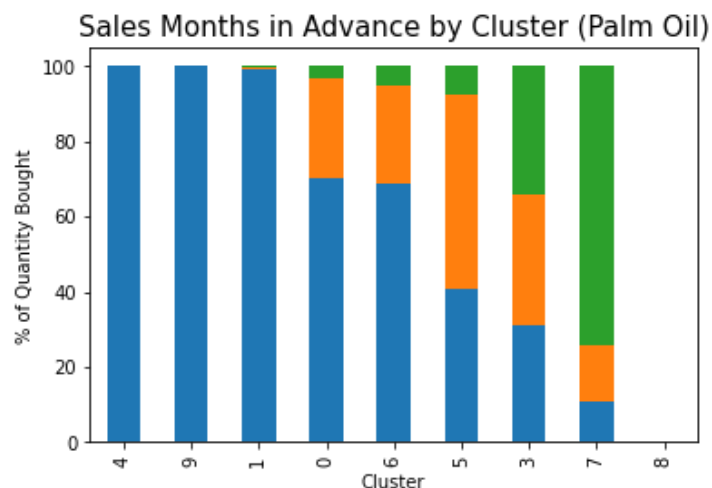
- **Majority of customers are short-term buyers** who predominantly buy contracts for delivery within 6 months or less
- Distinct group of **long to mid-term buyers** who consistently buy contracts for delivery after 6 months
- Segmentation holds across all oils

% of customers	Transaction Horizon		
	Short	Mid	Long
Palm	78.7	11.9	9.3
Sunflower	79.2	12.0	8.8
Rapeseed	59.6	23.1	17.4
Palm Kernel	84.2	13.6	2.2
Coconut	53.8	38.9	7.3
Soy	83.8	11.9	4.3

Short-term: <25% of sales for >6 months ahead of delivery

Mid term: <50% of sales made for >6 months ahead of delivery

Long term: >50% of sales made for >6 months ahead of delivery



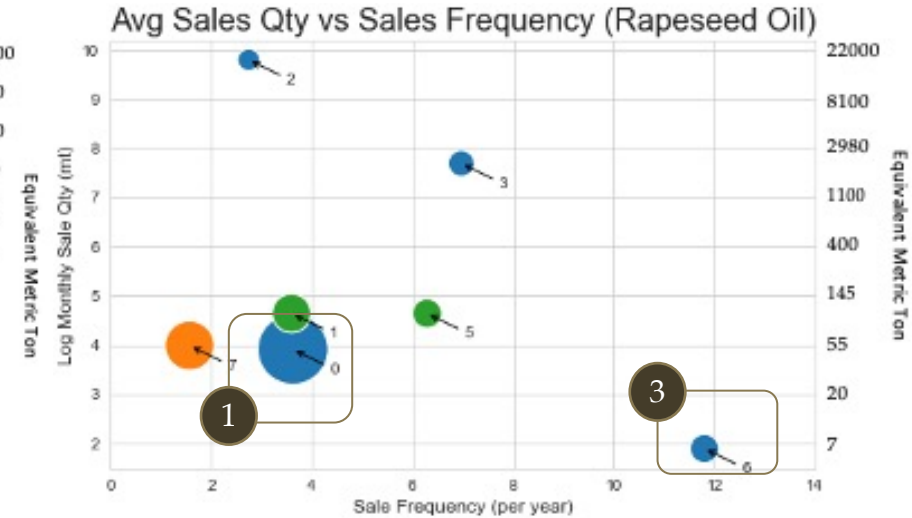
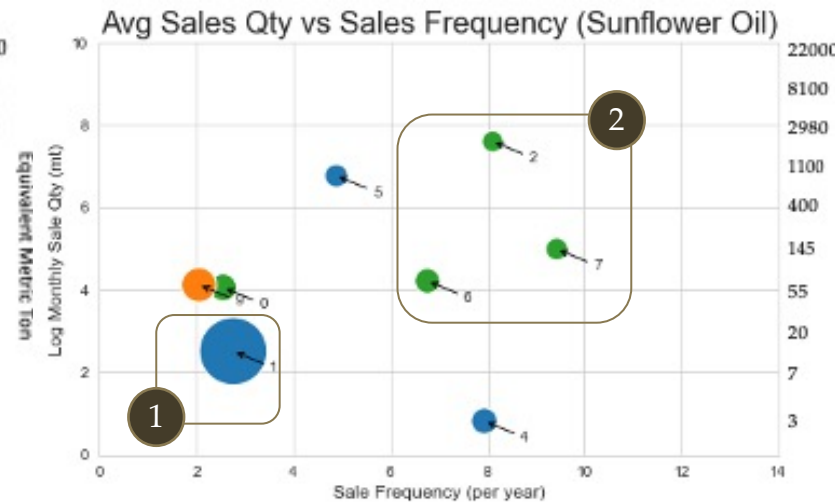
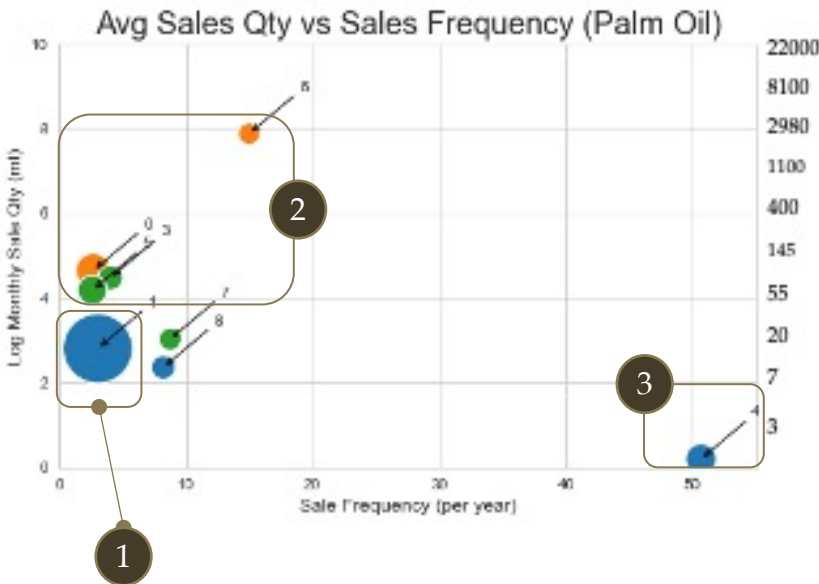
Clustering Results: Various groups based on sale frequency and quantum

- 1 Largest cluster of customers typically buys **infrequently** and in **lower quantities** each time (<4x a year, <20 mt/trnx)
- 2 Long-to-mid terms customer clusters tend to make sales more frequently or have larger average transaction size
- 3 Distinct higher-frequency, short-term sales customers for certain oils

**Size of cluster corresponds to number of observations in each cluster*

x-axis = Yearly Sales Frequency

y-axis = Log Monthly Sales Quantity (metric tons)



Full results in Appendix 2 & 3

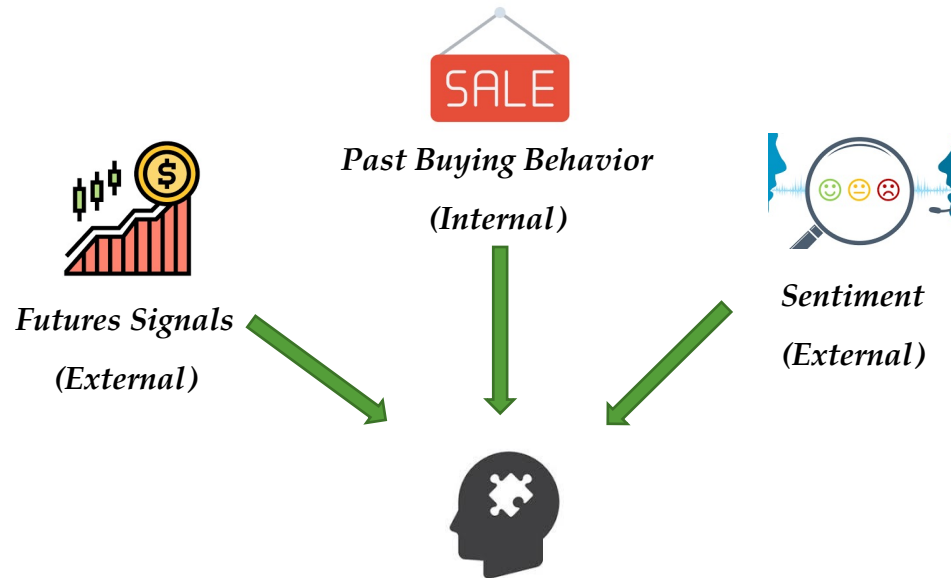


Part 2.1: Customer Transaction Forecasting

Using classification models to predict a customer sale event & identify sale triggers

Customer Transaction Forecasting Motivation

How should we think about sales forecasting?



CUSTOMER BEHAVIOR

2.1 Will customers have a sale in X month? (Binary)

2.2 What quantity will be sold, given sale = 1? (Continuous)

What are the potential applications?



- Enable ongoing advance hedging of positions in anticipation of customer sales
- Understand the internal and external factors that help to predict sales, to enable proactive monitoring

Data Preprocessing: Remove non-sales transactions, create customer features on a monthly level and set up monthly binary sales data over 2018-2021

Original Customer Transaction Dataset

(~450,000 rows, 6 oils)

1

Remove non-sales transactions

Below transactions are removed from the dataset

- $\text{monthly_pricing_quantity} \leq 0$ (17.6% of dataset)
- $\text{fixed_priced_per_mt} \leq 0$ (20.6% of dataset)
- $\text{fixed_priced_per_mt} \geq 4000$ (2.9% of dataset)
- $\text{sales_or_purchase tag} = \text{'purchase'}$ (1.6% of dataset)
- $\text{contract_line_status} = \text{'error'}$ (<0.2% of dataset)

2

Set up customer data on monthly level

- Set up dependent variable on a monthly level
- Compute all customer features on a monthly level
- Reduce noise from daily / weekly grain

3

Join external features to dataset

- Join relevant **futures** and **sentiment** data for the same day
- Features are aggregated by month, e.g. sentiment is averaged over entire month in question

4

Create lagged variables for prediction

- For time horizon in question, variables are lagged by appropriate interval. E.g. If predicting 3 months in advance, only features from 3 months before can be used.

Step 2 Example: If customer B3120 only bought in 1 month in 2018...

Counter Party Code	Contract Issue Date	... features
B3120	5/16/2018	...
B3120	5/10/2018	...

Counter Party Code	Month	Sale (D.V.)
...
B3120	4	0
B3120	5	1
B3120	6	0
B3120	7	0
...

Dependent variable for classification is sale or no sale in month

Understanding the Data

Prediction Dataset by Oil

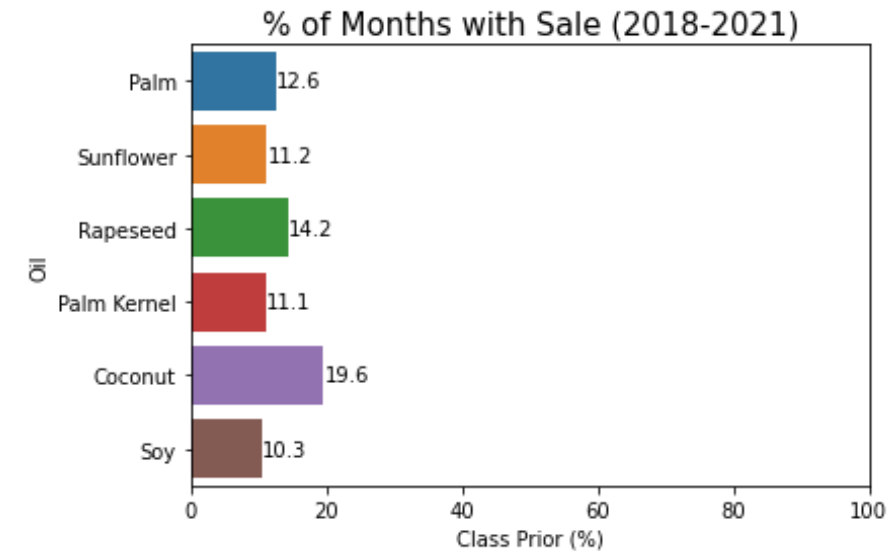
(over 2018 to 2021, 4 years)

1. Imbalanced class problem

- Month with sale made are <20% of dataset for all oils
- Imbalanced data handling methods needed

2. Palm, Sunflower and Rapeseed Oil are the top-3 most transacted oils

- Models built and tuned for all oils
- In-depth results discussed in the context of the above 3 oils



Oil	Unique Customers	Total Rows*	Sale	% of Months w/ Sale (class prior)
Palm	1,040	49,920	6,292	12.6
Sunflower	1,109	53,232	5,954	11.2
Rapeseed	663	31,824	4,528	14.2
Soy	452	21,696	2,245	10.3
Palm Kernel	400	19,200	2,135	11.1
Coconut	289	13,872	2,713	19.6

*Total Rows = Customers * 48 months (4 years of data) created

Sales Forecasting Approach

Built separate models for each oil

- Separate models instead of one unified (multiclass) model
- Different best model and best parameters for each oil
- Different futures price and sentiment signals factored in



Compared and tuned 3 classifiers to select best model

- Test 3 different approaches: Logistic Regression, RandomForest and LightGBM
- Allows us to test a range of approaches (linear, bagged and boosted)



Implemented imbalanced class handling in data pipeline

- Class weight parameters tuned for all three algorithms, resulting in increase of F1 score of ~10% vs equal class weights
- SMOTE-Tomek under + oversampling tested but F1 improvement equivalent to class weight parameter tuning



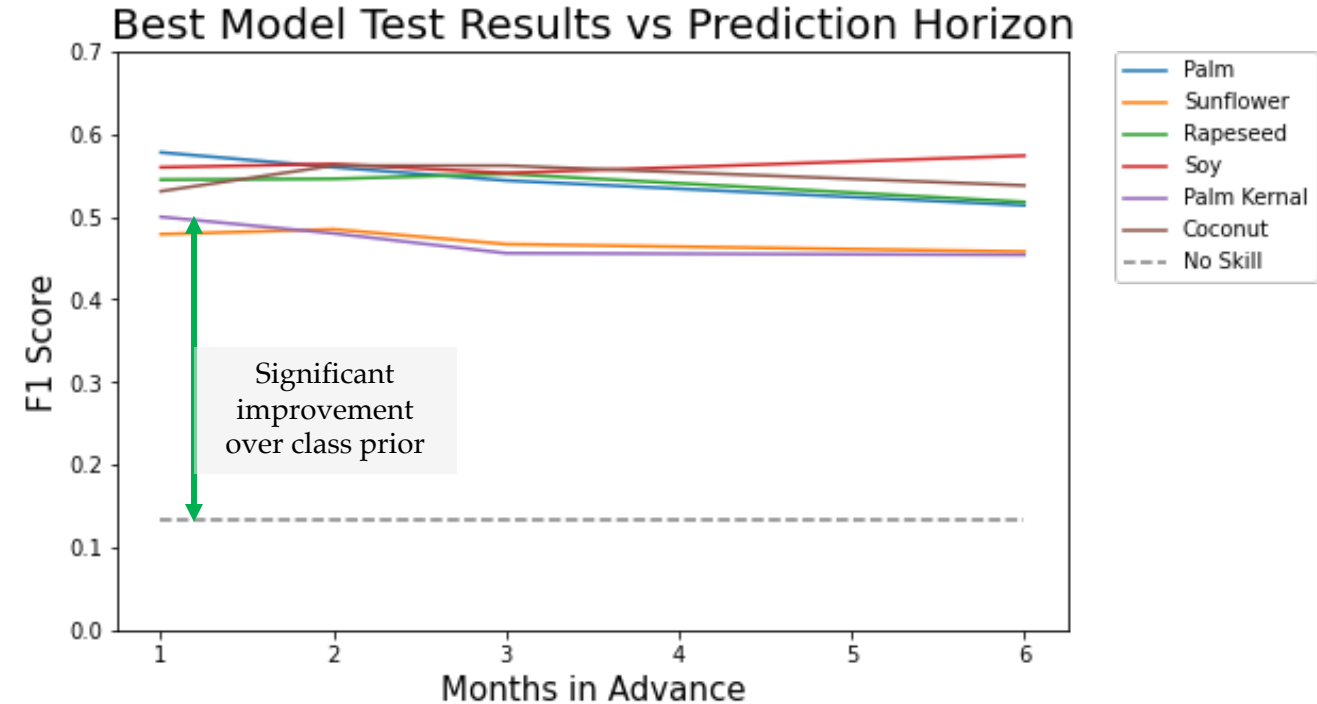
Tested model robustness over short, mid and long term

- Short term: Predicting sale event 1 to 2 months in advance
- Mid term: Predicting sale event 3 months in advance
- Long term: Predicting sale event 6 months in advance



Results: Up to 58% F1 in forecasting customer purchase in advance, ~50% F1 robust up to 6 months in advance, more than 3x class prior

- **Stable model performance** across short term (1, 2 months), mid term (3 months) and long term (6 months) on unseen data
- **Recent periods have strongest performance** i.e. 1 month F1 is typically highest, with slight drop off towards 6 months in advance
- **Significant improvement over avg class prior** (13% vs > 50% F1), providing incremental information vs guesswork
- Best model varies for each oil, but **bagged / boosted model generally outperforms** linear classifier



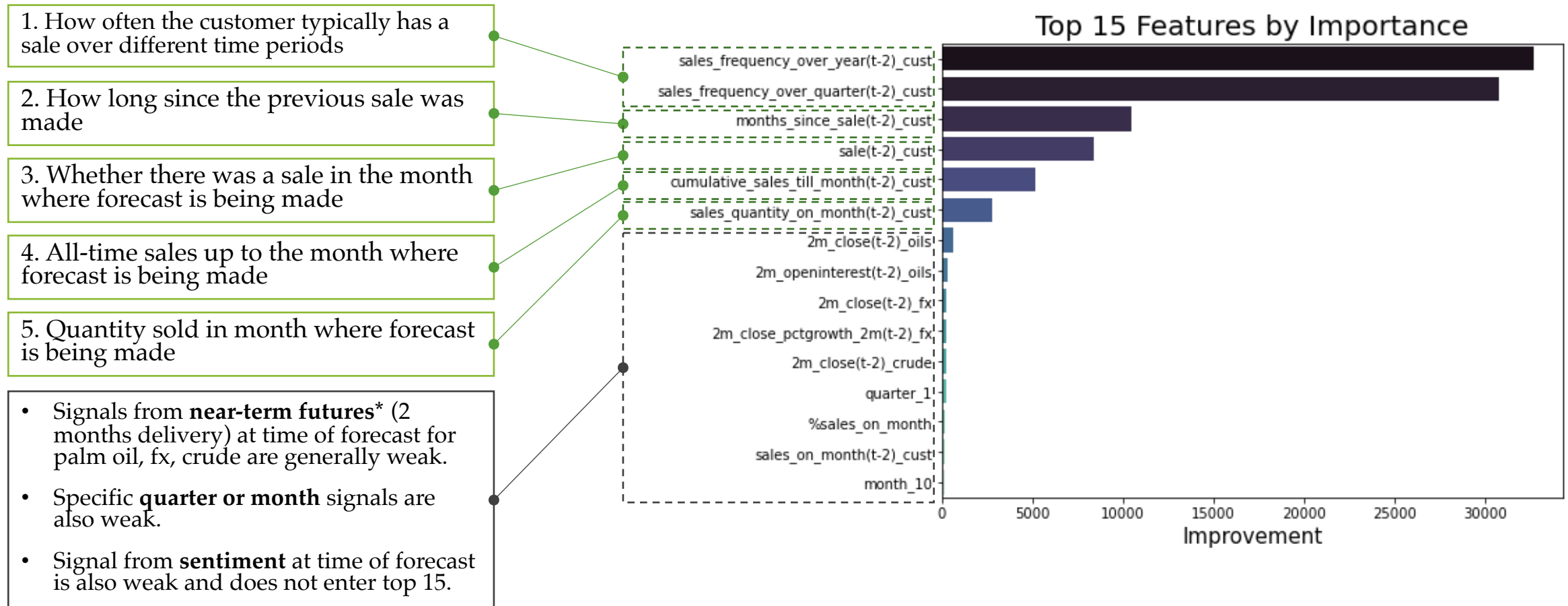
Oil	Palm	Sunflower	Rapeseed	Palm Kernal	Coconut	Soy
Best Model	LGBM	LGBM	LGBM	RF	LGBM	RF
Best Test F1	0.58 (1M)	0.49 (2M)	0.55 (3M)	0.50 (1M)	0.56 (3M)	0.57 (6M)
Best CV F1	0.57 (1M)	0.52 (1M)	0.56 (2M)	0.49 (1M)	0.56 (1M)	0.58 (1M)



Full results in Appendix 4

Sale Signals: Customers' past purchase behavior is the strongest predictor of future purchase (Palm Oil)

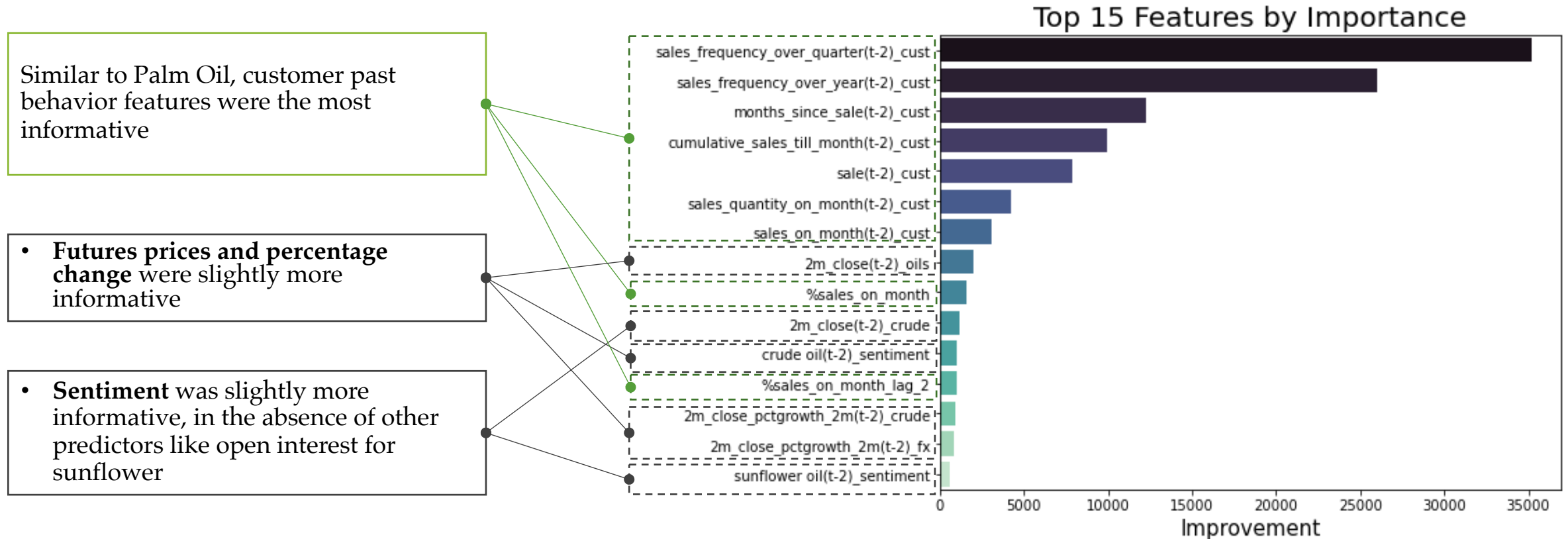
- Feature importance extracted from best model for **PALM OIL** (LightGBM), forecasting **2 months in advance**
- Features ranked by improvement which is the **cumulative reduction in logloss** (classification error) over all splits using feature i.e. how 'informative' the feature was in helping to make the correct classification




*Various futures time lags were tried (e.g. 1, 3, 6 months delivery) along with multiple combinations (e.g. 3 and 6, 2 and 4 etc.) with similar results due to high feature correlation.

Sale Signals: Customers' past purchase behavior is the strongest predictor of future purchase (Sunflower Oil)

- Feature importance extracted from best model for **SUNFLOWER OIL** (LightGBM), forecasting **2 months in advance**
- Trend generally holds for all other oils as well



Results for other oils in Appendix 5

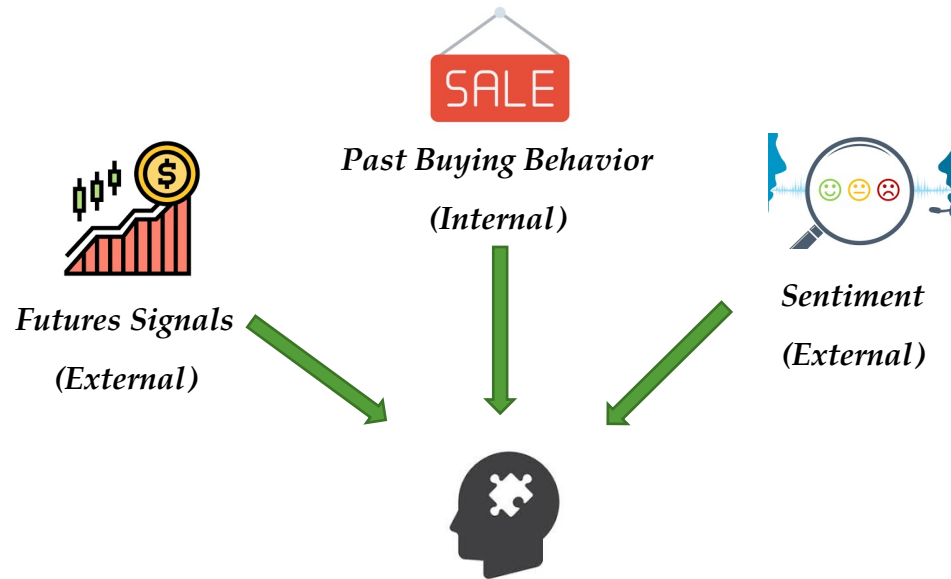


Part 2.2: Customer Purchase Amount Forecasting

Using linear regression models and feature selection algorithms to predict customer purchasing behavior

Customer Transaction Forecasting Motivation

How should we think about sales forecasting?



CUSTOMER BEHAVIOR

2.1 Will customers have a sale in X month? (Binary)

2.2 What quantity will be sold, given sale = 1? (Continuous)

What are the potential applications?



- Enable ongoing advance hedging of positions in anticipation of customer sales
- Understand the internal and external factors that help to predict sales, to enable proactive monitoring

Data Preprocessing: Remove non-sales transactions, create customer features, merge sentiment analysis, and add pricing data over 2018-2021

Original Customer Transaction Dataset

(~450,000 rows, 6 oils)

1

Remove non-sales transactions

Below transactions may be for recording or clearing purposes only

- $\text{monthly_pricing_quantity} \leq 0$ (17.6% of dataset)
- $\text{fixed_priced_per_mt} \leq 0$ (20.6% of dataset)
- $\text{fixed_priced_per_mt} \geq 4000$ (2.9% of dataset)
- $\text{sales_or_purchase tag} = \text{'purchase'}$ (1.6% of dataset)
- $\text{contract_line_status} = \text{'error'}$ (<0.2% of dataset)

2

Set up customer data on monthly level

- Set up dependent variable on a monthly level
- Compute all customer features on a monthly level
- Reduce noise from daily / weekly grain

3

Join external features to dataset

- Join relevant **futures** and **sentiment** data for the same day
- Features are aggregated by month, e.g. sentiment is averaged over entire month in question

4

Create lagged variables for prediction

- For time horizon in question, variables are lagged by appropriate interval. E.g. If predicting 3 months in advance, only features from 3 months before can be used.

5

Data Cleaning

- **Remove sales == 0 from individual oil set**
- Reduce skewness of individual features by log transformation

Final Dataset (~86600 rows, 6 oils)

Dependent variable is **monthly_pricing_quantity** aggregated by customer by month for each sale event recorded

Oil/Lag	Total Row	Sale	% Sales only	Oil/Lag	Total Row	Sale	% Sales only
Coconut 1m	13519	2672	20%	Rapeseed 1m	29172	4257	15%
Coconut 2m	13294	2626	20%	Rapeseed 2m	28509	4188	15%
Coconut 3m	13005	2582	20%	Rapeseed 3m	27846	4099	15%
Coconut 6m	12138	2422	20%	Rapeseed 6m	25857	3832	15%
Palm Kernel 1m	18000	2026	11%	Soy 1m	20340	2153	11%
Palm Kernel 2m	17600	1979	11%	Soy 2m	19888	2115	11%
Palm Kernel 3m	17200	1936	11%	Soy 3m	19436	2060	11%
Palm Kernel 6m	16000	1817	11%	Soy 6m	18080	1901	11%
Palm 1m	46800	5932	13%	Sunflower 1m	49905	5597	11%
Palm 2m	45760	5798	13%	Sunflower 2m	48796	5477	11%
Palm 3m	44720	5677	13%	Sunflower 3m	47687	5335	11%
Palm 6m	41600	5238	13%	Sunflower 6m	44360	4904	11%

Customer Sale Amount Forecasting Approach

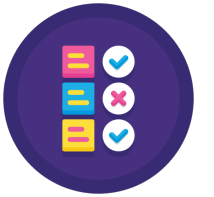
Built separate models for each oil and time-lag

- Separate regression models for each oil (6 oils) and each time-lag (4 time-lags), giving a total of 24 initial models.
- Different futures price, customer features and sentiment signals factored in.
- Linear instead of non-linear regression model chosen to ensure explainability. Performance of both models was comparable.



Reduced dimensionality with feature selection and regularization

- To reduce risk of multicollinearity, Max Relevance – Min Redundancy feature selection algorithm and Lasso Regularization were implemented.
- Best Lasso Regularization parameters were selected to boost performance of every model.
- The model with highest R-squared is selected for each oil.



Tested result robustness with statistical tests

- Check for statistical significance of all selected variables
- Remove insignificant variables and do a final backward feature selection to create a robust final model for each oil.



Results: Up to 41.0% R-Squared in forecasting customer sale qty and finding significant variables, increase of 6.2% due to feature selection and regularization

Table summary of best performing models for each oil type

Oil Type	Palm	Sunflower	Rapeseed	Palm Kernel	Coconut	Soy
Best Adjusted R-Squared	0.301	0.162	0.292	0.410	0.221	0.273
Best Time-Lag	3 months	1 month	1 month	1 month	3 months	1 month
No. of Features	7	6	6	7	4	3
Examples of Feature Selected	Sales Quantity, Commodity Sentiment Analysis, Monthly Sales	Sales Quantity, Petroleum Sentiment Analysis, Rapeseed Oil Sentiment Analysis	Sales Quantity, Energy Sentiment Analysis, Petroleum Sentiment Analysis	Sales Quantity, Corn Sentiment Analysis, Sunflower Oil Sentiment Analysis	Sales Quantity, Petroleum Sentiment Analysis, Sales Frequency	Sales Quantity, Sales Frequency, Monthly Sales

- **All 6 oil types have robust regression models** whereby all independent variables are statistically significant and are free from multicollinearity.
- **Coefficient of selected features are significant and meaningful (value > 0).** This gives us good insights into features that could affect purchasing quantity.
- Time-lag and regularization parameters vary for each oil, but models with **shorter time-lag (1 to 3 months) seem to outperform.** Quantity could be more influenced by more recent patterns.

Customer Sales Quantity Forecast (Palm Oil): Variety of internal and external signals contribute to amount sold

- The best performing model for **palm oil** was the one with **3-month feature lag** (i.e. prediction 3 months in advance)
- **7 features** were selected to build a robust model with **0.301 Adjusted R-Squared**.

CUSTOMER PAST BEHAVIOR (INTERNAL SIGNAL)

- **Sales Qty:** Larger past sales qty indicates larger future sales
- **Count of Sales:** Large number of past sales events indicates lower quantity of current sales, perhaps due to stock-up
- **Cumulative Sales:** All-time large purchasers tend to make larger sales

FUTURES (EXTERNAL SIGNAL)

- **Oil Open Interest:** Increasing open market interest also translates to higher purchase amounts
- **Crude Closing Price:** Crude is a general indicator of economic activity, hence increasing crude prices to a small extent indicates higher sales amounts

SENTIMENT (EXTERNAL SIGNAL)

- **Commodity:** In general, as sentiment tilts towards falling prices (i.e. positive sentiment), sales quantity reduces as customers may choose to close positions / go without contract
- **Soy Oil:** As soy and palm oils are substitutes in some applications, sentiment from soy oil is particularly significant

Selected Features	Coefficients
Constant	1.8579 (0.119)***
Log Sales Quantity on Month (t-3)	0.4157 (0.019)***
Log Sales on Month (t-3) <i>i.e. count of sales in month</i>	-0.696 (0.047)***
Log Cumulative Sales till Month (t-3)	0.3647 (0.012)***
Palm Oil Open Interest % Growth, 2M Delivery (t-3)	0.1567 (0.072)**
Crude Closing Price, 2M Delivery (t-3)	0.0104 (0.002)***
Commodity Sentiment (t-3)	-1.5592 (0.433)***
Soy Oil Sentiment (t-3)	-0.3069 (0.098)***

Notes: Four futures time lags were tried (e.g. 1, 2, 3, 6 months advance prediction) along with multiple regularization parameters, the best performing model was chosen.

*** Significant at $p < 0.01$ | ** Significant at $p < 0.05$

Customer Sales Quantity Forecast (Rapeseed Oil): Variety of internal and external signals contribute to amount sold

- The best performing model for **rapeseed oil** was the one with **1-month feature lag** (i.e. prediction 1 months in advance)
- **6 features** were adopted to build a robust model with **0.292 Adjusted R-Squared**.

CUSTOMER PAST BEHAVIOR (INTERNAL SIGNAL)

Similar directional relationships for **sales qty** and **count of sales** found as in palm oil.

- **Sales Frequency** (yearly basis): Customers who make sales more frequently also tend to buy more
- **% Sales on Month**: If customers are more likely to transact last month, they tend to buy less in current month

SENTIMENT (EXTERNAL SIGNAL)

- **Energy**: As sentiment tilts towards falling energy prices (i.e. positive sentiment), sales quantity increases as well. This could be because customers' other costs of production decrease and they hence increase production and in turn sales quantity of oils

FUTURES (EXTERNAL SIGNAL)

- No futures prices or volume movements were significant for rapeseed oil

Selected Features	Coefficients
Constant	3.9403 (0.053)
Log Sales Quantity on Month (t-1)	0.5127 (0.017)***
Log Sales on Month (t-1) <i>i.e. count of sales in month</i>	-1.1952 (0.053)***
Log Sales Frequency on Year (t-1)	0.7713 (0.038)***
Log % Sales on Month (t-1)	-6.1302 (0.78)***
Energy Sentiment (t-1)	0.2390 (0.11)**
Gas Sentiment (t-1)	0.1757 (0.132)

Notes: Four futures time lags were tried (e.g. 1, 2, 3, 6 months advance prediction) along with multiple regularization parameters, the best performing model was chosen.

*** Significant at $p < 0.01$ | ** Significant at $p < 0.05$



Results for other oils in Appendix 6



Part 3: Determinants of Futures Prices

Suggested approach to predict
determinants of future prices using linear
regression models and feature selection
algorithms

Data Preprocessing: Remove Merge sentiment analysis, and lag pricing data over 2018-2021 periods

Original Futures Dataset

(~76,000 rows, 5 oils)

1

Set up futures price data on a monthly basis

- Dependent variable: Average monthly futures closing price by oil
- Monthly level prediction** was chosen as we aim to reduce noise from weekly or daily fluctuations

2

Join relevant features to dataset

- Join relevant **sentiment** data for the same month
- Join relevant crude and other oils closing prices, open interest and volume for the same month where available

3

Create lagged variables for prediction

- For time horizon in question, variables are lagged by appropriate interval. E.g. If predicting 3 months in advance, only features from 3 months before can be used.

4

Repeat Steps 1-3 for different delivery horizon

- Get similar data set up for each type of future e.g. delivery 1 month from forecast month, 2 months from forecast month and 6 months from forecast month

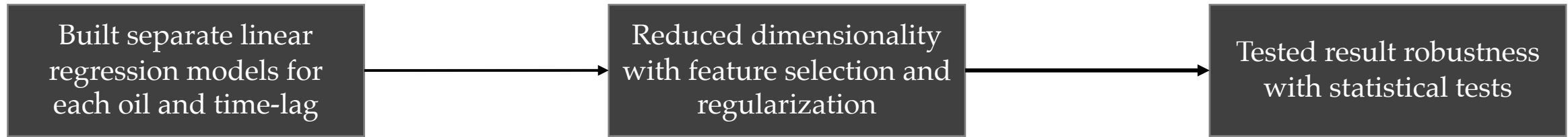
Final Dataset by Future

(~860 rows, 5 oils)

Oil/Lag	Total Row	Oil/Lag	Total Row
Palm Kernel 1m	45	Soy 1m	45
Palm Kernel 2m	44	Soy 2m	44
Palm Kernel 3m	43	Soy 3m	43
Palm Kernel 6m	40	Soy 6m	40
Palm 1m	45	Sunflower 1m	45
Palm 2m	44	Sunflower 2m	44
Palm 3m	43	Sunflower 3m	43
Palm 6m	40	Sunflower 6m	40
Rapeseed 1m	44		
Rapeseed 2m	43		
Rapeseed 3m	42		
Rapeseed 6m	39		

Price Prediction Challenges: Purely using pricing, volume, date and sentiment data is insufficient to produce a meaningful signal for futures price prediction

Forecasting Approach and Results



- Our suggested approach is similar to that of forecasting customer purchase amount as the intent is to find the best features which essentially require us to find **causality**.
- However, overall r-squared across all 24 models explored was **effectively 0.0**, suggesting that futures prices are generally unpredictable or that including past future prices / sentiment is insufficient.

Challenges

- Insufficient covariates to **produce generalizable results** because only purchase and sentiment analysis data was included in models.
- Future exploration could include data that could be relevant to oil prices such as shipping patterns or global supply level, which are sources for information not available to us now.



Conclusion

Business application and suggested improvements to models

In Summary

CUSTOMER SEGMENTATION



TRANSACT/ NOT TRANSACT



PRICE ATTRIBUTION



BUSINESS APPLICATION

- **Initial Customer Profiling:**
Use to bucket newly onboarded customers based on their industry, production capacity, etc. as Cargill would have no transactional data available for these entities.
- **Customized Client Support:**
Provide specialized support to customers e.g. specialists in short, mid and long-term hedging or dedicated support for large customers
- **Transaction Trigger:**
Better understanding of determinants influencing sales activity and volumes for proactive monitoring
- **Proactive Hedging:**
Allow Cargill to manage positions in advance of actual customer activity, potentially increasing ability to hedge or give said advice to clients as a packaged value-add service
- **Not Ready for Production Use:**
Requires more data and tuning before use in production.

SUGGESTED IMPROVEMENTS

- **Data Enrichment:**
Train model with more complete customer information (e.g., industry, market cap, etc.)
- **Decision Support System:**
Link to decision support system prompting desk or customer to buy / sell
- **Explore Internal Trades:**
Potential off market transactions by pre-matching sell and demand side clients, translating to lower transaction costs.
- **Data Enrichment:**
Retrain models with more data as insufficient samples were insufficient making the model biased.



Appendix

I. Data Dictionary

Category	Feature Name	Explanation	Clustering	Classification	Regression
Date	month	Month of the year			
Date	quarter	Quarter of the year (1 = Jan to Mar, 2 = Apr to Jun, 3 = Jul to Sep, 4 = Oct to Dec)			
Customer	sale	Binary value; Whether a sale was made on selected month for selected customer		D.V.	
Customer	sales_on_month	Count of the number of sales made on selected month for selected customer			
Customer	sales_quantity_on_month	Total oil quantity sold on selected month for selected customer			D.V.
Customer	cumulative_sales_per_month	Cumulative number of sales made till selected month for selected customer			
Customer	sales_frequency	The average number of sales per month for selected customer			
Customer	sales_frequency_over_quarter	The average number of sales per month for selected customer calculated over the past quarter (3 months)			
Customer	sales_frequency_over_year	The average number of sales per month for selected customer calculated over the past year			
Customer	months_since_sale	Number of months since last sale for selected customer			
Customer	%sales_on_month	Percentage of sales made on selected month compared to total sales made for a selected customer			
Customer	sell_qty_%0-5+ Months	Percent of total quantity sold for delivery (0,6) months from date of trade			
Customer	sell_qty_%6-12 Months	Percent of total quantity sold for delivery (6,12] months from date of trade			
Customer	sell_qty_%12+ Months	Percent of total quantity sold for delivery more than 12 months from date of trade			
Customer	sell_qty_total	Total oil quantity bought			
Futures	close	Average closing price of futures (fx, edible oil, crude) for the month from GetHistory API (where available, else Reuters)			
Futures	close_pctgrowth	Percentage growth of closing price. Suffix refers to number of preceding months over which growth is computed.			
Futures	openinterest	Average open interest of futures (fx, edible oil, crude) for the month from GetHistory API (where available)			
Futures	openinterest_pctgrowth	Percentage growth of open interest. Suffix refers to number of preceding months over which growth is computed.			
Futures	volume	Sum of total volume traded (fx, edible oil, crude) for the month from GetHistory API (where available)			
Futures	volume_pctgrowth	Percentage growth of volume. Suffix refers to number of preceding months over which growth is computed.			
Sentiment	crude oil	The average sentiment value per month for selected commodity - Keywords [oil crude clz clf hoz hof]			
Sentiment	crude oil_Count	Total sentiment count on selected month for selected commodity - Keywords [oil crude clz clf hoz hof]			
Sentiment	palm oil	The average sentiment value per month for selected commodity - Keywords [palm]			
Sentiment	palm oil_Count	Total sentiment count on selected month for selected commodity - Keywords [palm]			
Sentiment	soy	The average sentiment value per month for selected commodity - Keywords [soy soybeans]			
Sentiment	soy_Count	Total sentiment count on selected month for selected commodity - Keywords [soy soybeans]			

Note: Prefix for futures prices refers to contract for delivery X months from current month e.g. 3m_close refers to closing price of contract for delivery in 3 months time.

Variable was used

I. Data Dictionary (2/2)

Category	Feature Name	Explanation	Clustering	Classification	Regression
Sentiment	rapeseed oil	The average sentiment value per month for selected commodity - Keywords [rapeseed]			
Sentiment	rapeseed oil_Count	Total sentiment count on selected month for selected commodity - Keywords [rapeseed]			
Sentiment	sunflower oil	The average sentiment value per month for selected commodity - Keywords [sunflower]			
Sentiment	sunflower oil_Count	Total sentiment count on selected month for selected commodity - Keywords [sunflower]			
Sentiment	coconut oil	The average sentiment value per month for selected commodity - Keywords [coconut]			
Sentiment	coconut oil_Count	Total sentiment count on selected month for selected commodity - Keywords [coconut]			
Sentiment	energy	The average sentiment value per month for selected commodity - keywords [energy energies]			
Sentiment	energy_Count	Total sentiment count on selected month for selected commodity - keywords [energy energies]			
Sentiment	corn	The average sentiment value per month for selected commodity - keywords [corn]			
Sentiment	corn_Count	Total sentiment count on selected month for selected commodity - keywords [corn]			
Sentiment	wheat	The average sentiment value per month for selected commodity - keywords [wheat]			
Sentiment	wheat_Count	Total sentiment count on selected month for selected commodity - keywords [wheat]			
Sentiment	gas	The average sentiment value per month for selected commodity - keywords [gas]			
Sentiment	gas_Count	Total sentiment count on selected month for selected commodity - keywords [gas]			
Sentiment	grains	The average sentiment value per month for selected commodity - keywords [grain crop bean coffee weather]			
Sentiment	grains_Count	Total sentiment count on selected month for selected commodity - keywords [grain crop bean coffee weather]			
Sentiment	ethanol	The average sentiment value per month for selected commodity - keywords [ethanol]			
Sentiment	ethanol_Count	Total sentiment count on selected month for selected commodity - keywords [ethanol]			
Sentiment	market	The average sentiment value per month for selected commodity - keywords [market]			
Sentiment	market_Count	Total sentiment count on selected month for selected commodity - keywords [market]			
Sentiment	gold	The average sentiment value per month for selected commodity - keywords [gold]			
Sentiment	gold_Count	Total sentiment count on selected month for selected commodity - keywords [gold]			
Sentiment	s&p	The average sentiment value per month for selected commodity - keywords [s&p dow]			
Sentiment	s&p_Count	Total sentiment count on selected month for selected commodity - keywords [s&p dow]			
Sentiment	commodity	The average sentiment value per month for selected commodity - keywords [commodity commodities]			
Sentiment	commodity_Count	Total sentiment count on selected month for selected commodity - keywords [commodity commodities]			
Sentiment	cattle	The average sentiment value per month for selected commodity - keywords [cattle beef]			
Sentiment	cattle_Count	Total sentiment count on selected month for selected commodity - keywords [cattle beef]			
Sentiment	silver	The average sentiment value per month for selected commodity - keywords [silver]			
Sentiment	silver_Count	Total sentiment count on selected month for selected commodity - keywords [silver]			
Sentiment	platinum	The average sentiment value per month for selected commodity - keywords [platinum]			
Sentiment	platinum_Count	Total sentiment count on selected month for selected commodity - keywords [platinum]			

Variable was used

2. Cluster Aggregate Features by Oil (1/2)

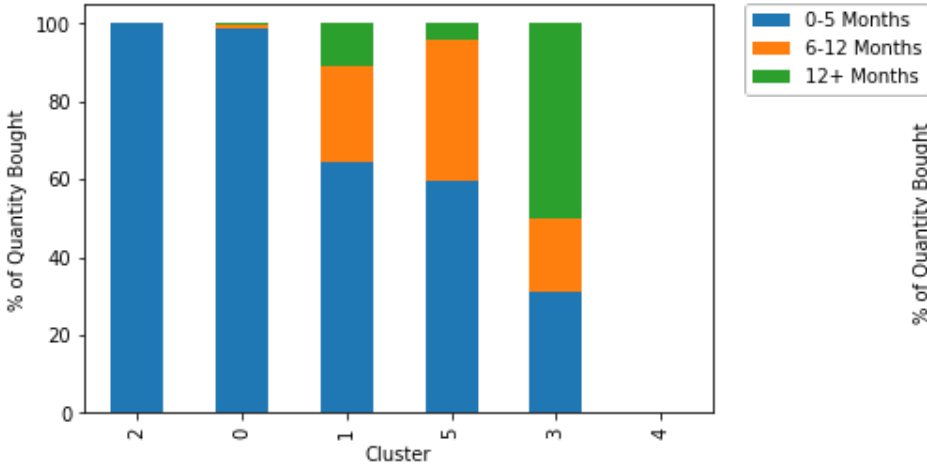
	Cluster	Customers	% of total	Average Yearly Sale Frequency	Avg Monthly Sales Qty (mt)	Log Avg Monthly Sales Qty	Sales Horizon
Palm	0	113	11.6	2.6	102.8	4.6	Mid
	1	665	68.2	3.0	15.5	2.8	Short
	3	24	2.5	4.0	86.6	4.5	Long
	4	72	7.4	50.7	0.2	0.2	Short
	5	61	6.3	2.6	64.6	4.2	Long
	6	3	0.3	15.0	2634.0	7.9	Mid
	7	6	0.6	8.7	19.6	3.0	Long
	8	18	1.8	8.2	9.6	2.4	Short
	9	12	1.2	129.0	0.1	0.1	Short
Sunflower	0	50	4.8	2.5	57.1	4.1	Long
	1	689	65.6	2.8	11.4	2.5	Short
	2	2	0.2	8.1	1999.9	7.6	Long
	3	12	1.1	162.0	10.5	2.4	Short
	4	40	3.8	7.9	1.3	0.8	Short
	5	16	1.5	4.9	870.0	6.8	Short
	6	32	3.0	6.8	67.1	4.2	Long
	7	9	0.9	9.4	146.1	5.0	Long
	8	75	7.1	57.6	0.5	0.4	Short
	9	126	12.0	2.0	60.9	4.1	Mid
Rapeseed	0	304	52.3	3.6	48.8	3.9	Short
	1	70	12.0	3.6	102.7	4.6	Long
	2	2	0.3	2.7	17866.5	9.8	Short
	3	15	2.6	7.0	2176.5	7.7	Short
	4	4	0.7	135.0	11.1	2.5	Long
	5	27	4.6	6.3	102.3	4.6	Long
	6	25	4.3	11.8	5.6	1.9	Short
	7	134	23.1	1.6	52.8	4.0	Mid

2. Cluster Aggregate Features by Oil (2/2)

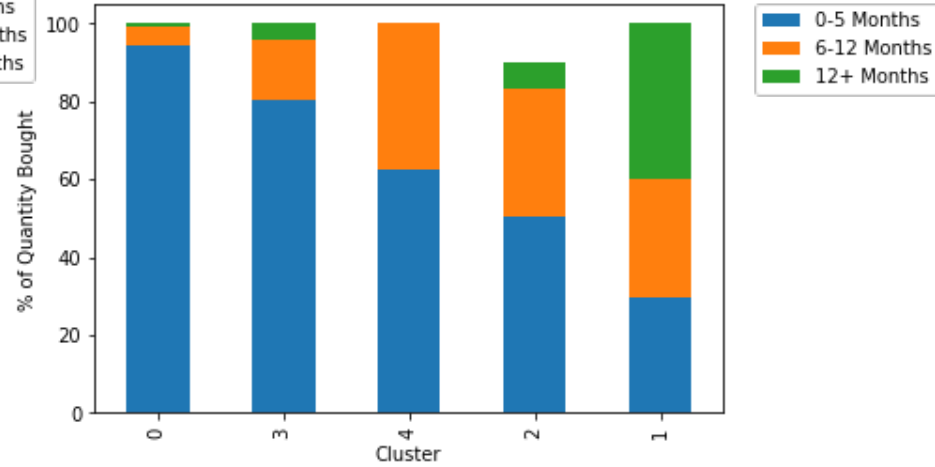
	Cluster	Customers	% of total	Average Yearly Sale Frequency	Avg Monthly Sales Qty (mt)	Log Avg Monthly Sales Qty	Sales Horizon
Palm Kernel	0	282	76.6	3.9	5.2	1.8	Short
	1	6	1.6	15.3	267.9	5.6	Mid
	2	20	5.4	99.6	0.1	0.1	Short
	3	8	2.2	4.6	7.2	2.1	Long
	4	8	2.2	6.3	0.1	0.1	Short
	5	44	12.0	3.2	22.4	3.2	Mid
Coconut	0	136	51.9	2.1	11.5	2.5	Short
	1	19	7.3	4.0	15.6	2.8	Long
	2	98	37.4	2.6	45.9	3.8	Mid
	3	5	1.9	17.5	498.9	6.2	Short
	4	4	1.5	78.0	2.2	1.1	Mid
Soy	0	274	69.2	2.6	44.8	3.8	Short
	1	47	11.9	3.0	32.0	3.5	Mid
	2	1	0.3	7.6	6272.8	8.7	Short
	3	10	2.5	3.1	10.7	2.5	Long
	4	9	2.3	77.3	3.5	1.5	Short
	5	48	12.1	7.6	7.0	2.1	Short
	6	7	1.8	33.7	5.2	1.8	Long

3. Cluster Results for Other Oils

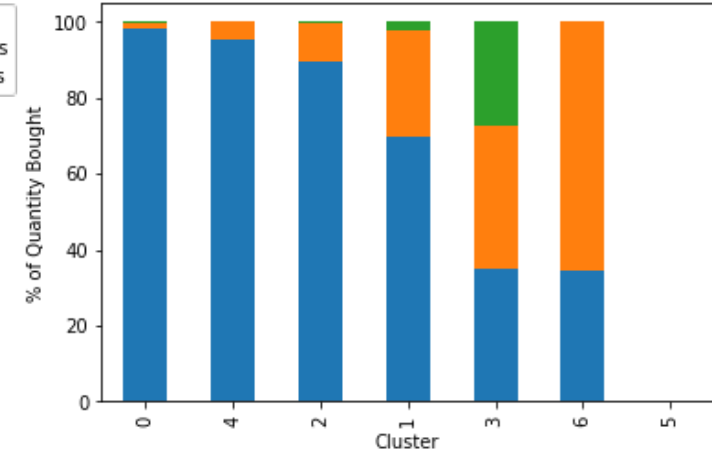
Sales Months in Advance by Cluster (Palm Kernal Oil)



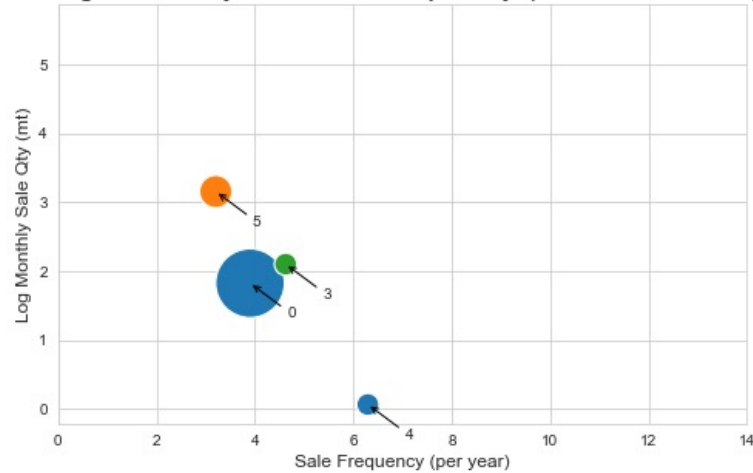
Sales Months in Advance by Cluster (Coconut Oil)



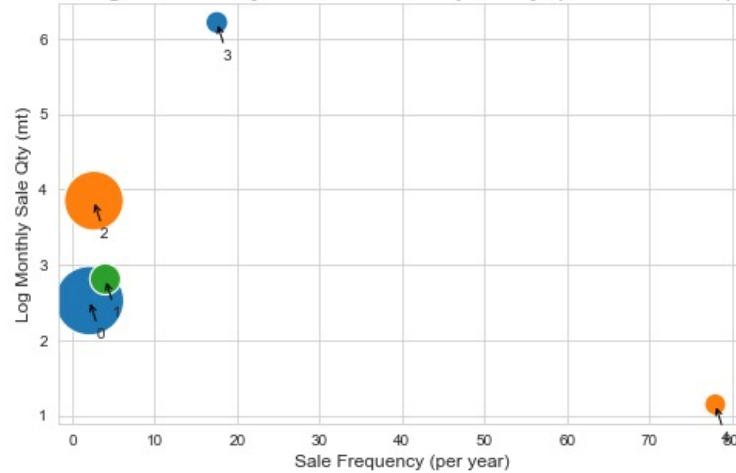
Sales Months in Advance by Cluster (Soy Oil)



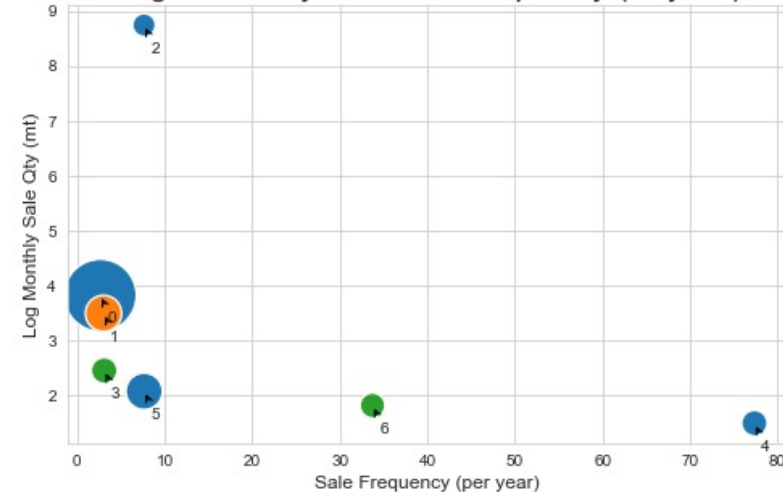
Avg Sales Qty vs Sales Frequency (Palm Kernal Oil)



Avg Sales Qty vs Sales Frequency (Coconut Oil)



Avg Sales Qty vs Sales Frequency (Soy Oil)

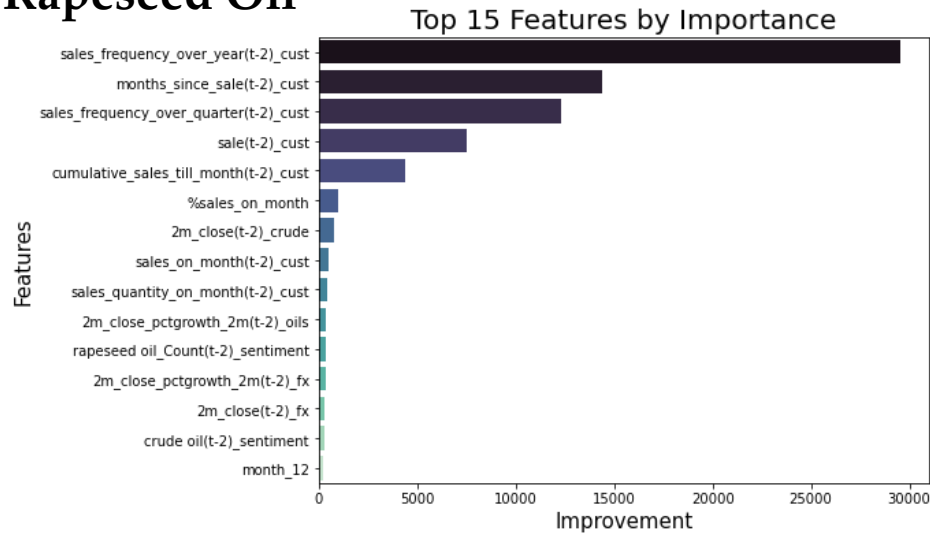


4. Classification Results by Model

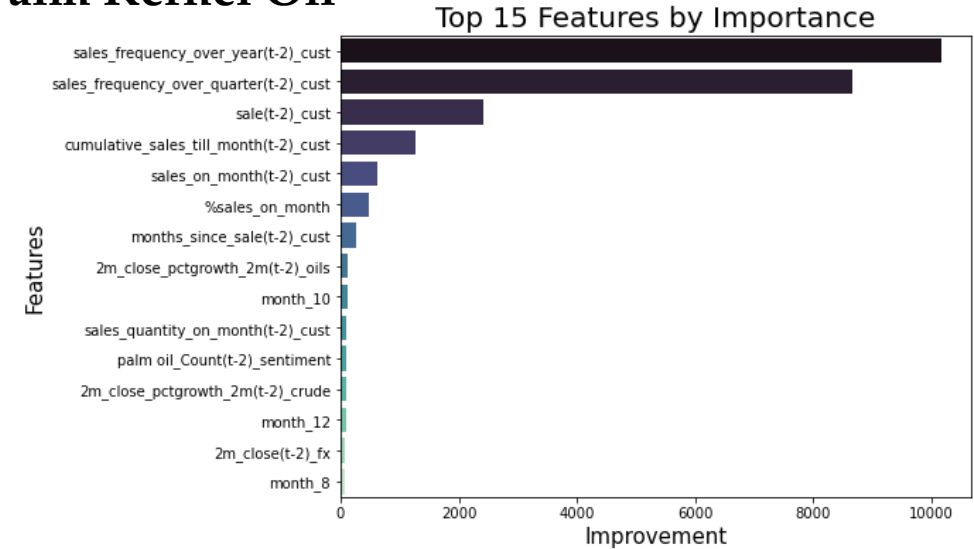
		Test Results (F1)				Cross Validation Results (F1)			
Oil	Model	<i>1 month</i>	<i>2 months</i>	<i>3 months</i>	<i>6 months</i>	<i>1 month</i>	<i>2 months</i>	<i>3 months</i>	<i>6 months</i>
Palm	Logistic Regression	0.541	0.535	0.528	0.501	0.545	0.535	0.529	0.503
	Random Forest	0.567	0.557	0.550	0.518	0.565	0.552	0.545	0.516
	LightGBM	0.578	0.560	0.544	0.514	0.565	0.559	0.547	0.521
Sunflower	Logistic Regression	0.456	0.472	0.458	0.437	0.492	0.481	0.462	0.434
	Random Forest	0.463	0.488	0.466	0.448	0.506	0.493	0.474	0.451
	LightGBM	0.479	0.485	0.467	0.458	0.515	0.492	0.473	0.450
Rapeseed	Logistic Regression	0.512	0.522	0.542	0.518	0.535	0.542	0.531	0.517
	Random Forest	0.550	0.535	0.544	0.515	0.554	0.547	0.538	0.523
	LightGBM	0.545	0.546	0.552	0.518	0.549	0.557	0.539	0.523
Palm Kernel	Logistic Regression	0.485	0.457	0.455	0.433	0.465	0.474	0.464	0.432
	Random Forest	0.500	0.480	0.456	0.454	0.494	0.477	0.480	0.454
	LightGBM	0.483	0.463	0.445	0.448	0.474	0.469	0.464	0.438
Coconut	Logistic Regression	0.515	0.563	0.554	0.552	0.544	0.543	0.538	0.535
	Random Forest	0.514	0.563	0.544	0.524	0.557	0.548	0.552	0.544
	LightGBM	0.531	0.562	0.562	0.538	0.560	0.551	0.544	0.552
Soy	Logistic Regression	0.508	0.500	0.514	0.503	0.539	0.516	0.510	0.489
	Random Forest	0.560	0.564	0.553	0.574	0.575	0.565	0.550	0.543
	LightGBM	0.554	0.542	0.575	0.551	0.572	0.553	0.544	0.529

5. Classification LGBM Feature Importance for Other Oils

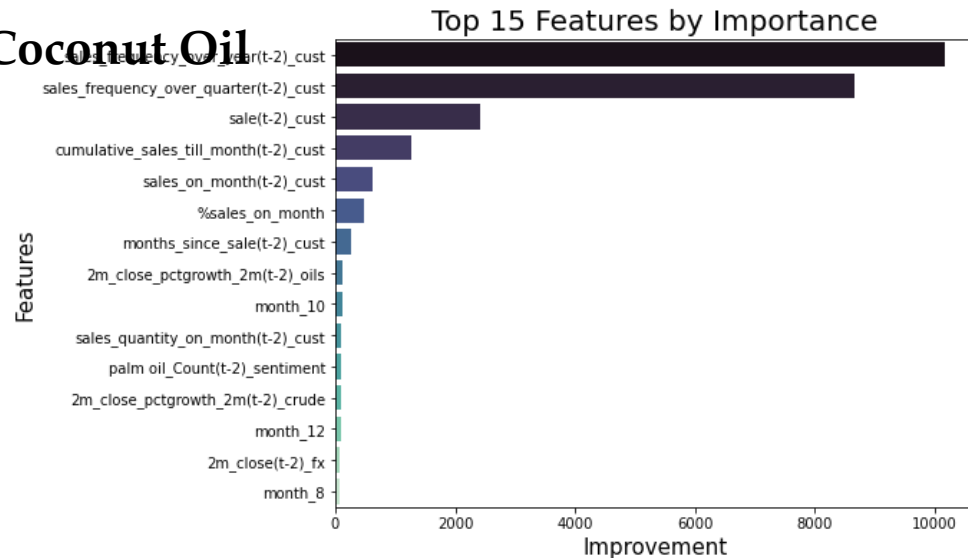
Rapeseed Oil



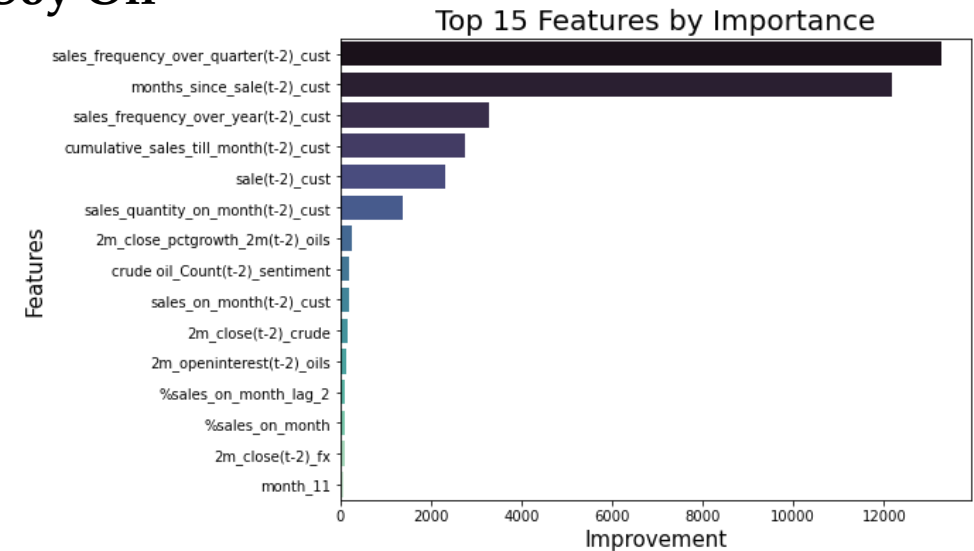
Palm Kernel Oil



Coconut Oil



Soy Oil



6. Regression Coefficients for Other Oils

Sunflower Oil

Selected Features	Coefficients
Constant	-1.4326 (0.797)
Log Sales Quantity on Month (t-1)	0.5462 (0.019)
Log Sales on Month (t-1)	-0.8438 (0.049)
Log % Sales on Month	-1.9515 (0.668)
Sunflower Oil Closing Price, 2M Delivery (t-1)	0.7561 (0.118)
Gas Sentiment (t-1)	0.1596 (0.133)
Rapeseed Oil Sentiment (t-1)	0.2003 (0.137)

Palm Kernel Oil

Selected Features	Coefficients
Constant	1.1009 (0.084)
Log Sales Quantity on Month (t-1)	0.583 (0.036)***
Log Sales on Month (t-1)	-1.0985 (0.089)***
Log Cumulative Sales till Month (t-1)	0.467 (0.046)***
Log Monthly Sales Frequency (t-1)	0.1935 (0.137)
Corn Sentiment (t-1)	-0.4138 (0.126)***
Sunflower Oil Sentiment (t-1)	-0.7852 (0.275)***
Cattle Sentiment Volume (t-1)	0.0212 (0.01)**

Coconut Oil

Selected Features	Coefficients
Constant	79.3419 (10.546)
Log Sales Quantity on Month (t-3)	0.5035 (0.026)***
Log Sales on Month (t-3)	-15.1628 (1.392)***
Log Monthly Sales Frequency (t-3)	30.2779 (2.036)***
Gas Sentiment (t-3)	102.6287 (34.852)***

Soy Oil

Selected Features	Coefficients
Constant	3.5537 (0.052)
Log Sales Quantity on Month (t-1)	0.4171 (0.022)
Log Sales on Month (t-1)	-0.8428 (0.07)
Log Monthly Sales Frequency (t-1)	0.722 (0.056)