Aditya Gohain
MSDS 451 - Financial Machine Learning
Week 03 - Programming Assignment 01
October 5th, 2025

Financial Machine Learning for Agricultural Commodities

Problem Description

Working as a drivetrain engineer for John Deere, I have had the opportunity to learn several insights into the highly cyclical nature of the agricultural heavy machinery sector. The cyclicality is typically influenced by various environmental factors such as global weather patterns, crop conditions, evolving diets, technological advancements, and commodity price fluctuations. Since this course presented me an opportunity to analyze any market, I wanted to understand the overall financial impact from price fluctuation of major commodities such as Corn, Soybeans, Cotton, and Wheat, and how that may impact revenue or product sales for companies in agricultural heavy machinery industry. My objective was to develop a machine learning model for predicting short-term price movements (upward, downward, or stable) using historical financial data of the futures for these commodities. This project focused on feature engineering, model building, and evaluation.

Data Preparation, Pipeline & Programming

I used historical daily futures price data between (200-2025) sourced using Yahoo! Finance for each of the four commodities. For each of the datasets, I created a pipeline for preprocessing and feature engineering by deriving the following features/indications:

- **Lagged prices** for 1–3 days or today's price, price 2 days ago, and price 3 days ago.
- Similarly, **tading volume lags** for 1-3 days
- **Exponential moving averages** using the lagged prices in order to mitigate leakage which acted as indicators for momentum.
- **Log returns**, by calculating the natural log of relative prices which was used as a regresstion target as well as for classifying.
- **Binary targets** where 1 indicated if the returns were positive while 0 indicated negative returns

Notes

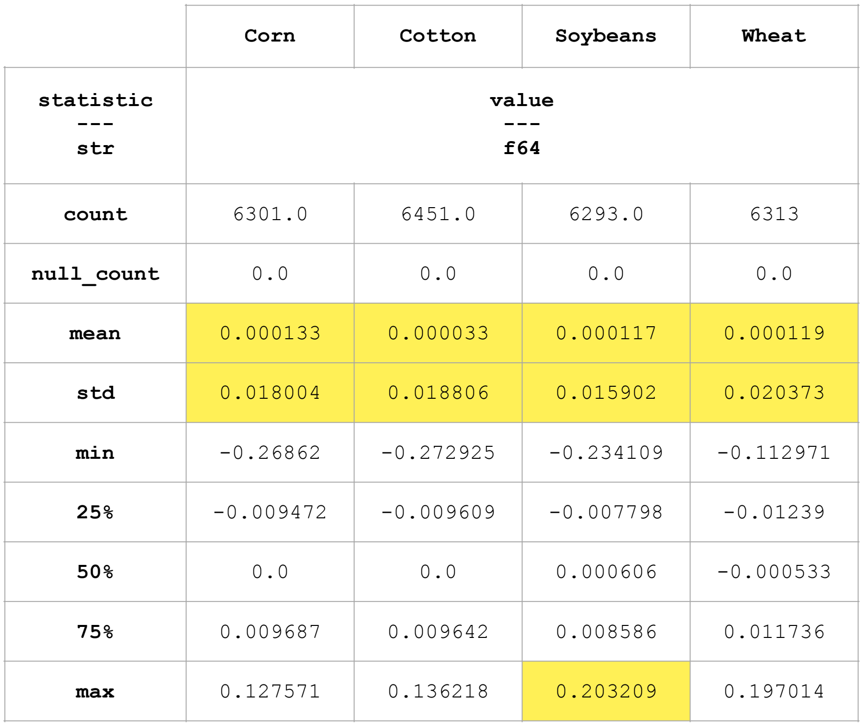
- In terms of the programming process, I removed any null values that got created from the lagging process and standardized the data prior to training the models.

- The analysis was primarily done using **Python** and I used the **Polars** and **Pandas** libraries to clean the data, transform tables, and compute the features. I also incorporated **scikit-learn** to preprocess, cross-valdate, and to perform logistical regression and derive evaluation metrics. Lastly, I used **Matplotlib** and **Seaborn** for creating time series vizualizations and heatmaps.

Research Design

Results and Insights for Agricultural Commodities

Distribution of Log Returns

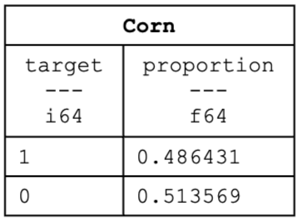


	Corn	Cotton	Soybeans	Wheat
statistic --- str	value --- f64			
count	6301.0	6451.0	6293.0	6313
null_count	0.0	0.0	0.0	0.0
mean	0.000133	0.000033	0.000117	0.000119
std	0.018004	0.018806	0.015902	0.020373
min	-0.26862	-0.272925	-0.234109	-0.112971
25%	-0.009472	-0.009609	-0.007798	-0.01239
50%	0.0	0.0	0.000606	-0.000533
75%	0.009687	0.009642	0.008586	0.011736
max	0.127571	0.136218	0.203209	0.197014

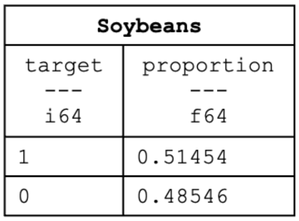
After I analyzed daily log returns across Corn, Soybeans, Wheat, and Cotton I saw several consistent patterns:

The mean daily return for all commodities was close to zero, this indicated to me a consistency accross the commodity markets and in a way showed indication of an efficient market behavior. In terms of volatility, looking at the standard deviation of returns also showed similarity across commodities, with averages being about 1.6–2.0%. Wheat interestingly displays a slightly higher volatility in comparison to Corn and Soybeans. Soybeans seem to stand out for having the single largest observed daily gain (max) of about 20%, which highlights their exposure to extreme price shocks. Looking at the minimum values, Corn and Cotton dropped to around -27% but Wheat dropped to -11%. The difference here is larger in magnitude than the positive gains (max) which relates to a negative skewness and higher downside risk related to agricultural commodities.

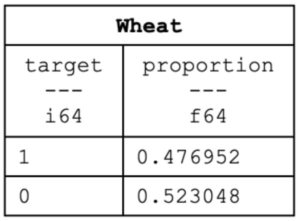
Target Class Distribution



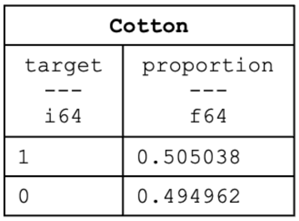
Corn	
target --- i64	proportion --- f64
1	0.486431
0	0.513569



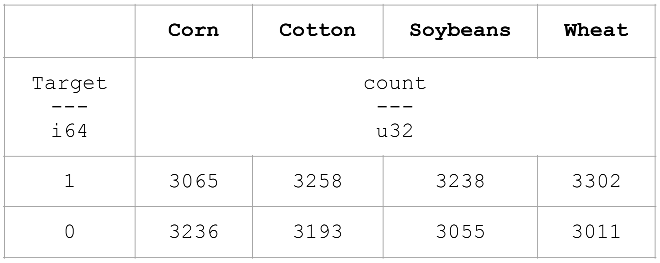
Soybeans	
target --- i64	proportion --- f64
1	0.51454
0	0.48546



Wheat	
target --- i64	proportion --- f64
1	0.476952
0	0.523048



Cotton	
target --- i64	proportion --- f64
1	0.505038
0	0.494962

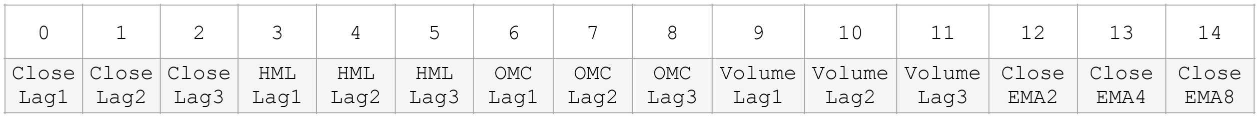


	Corn	Cotton	Soybeans	Wheat
Target --- i64	count --- u32			
1	3065	3258	3238	3302
0	3236	3193	3055	3011

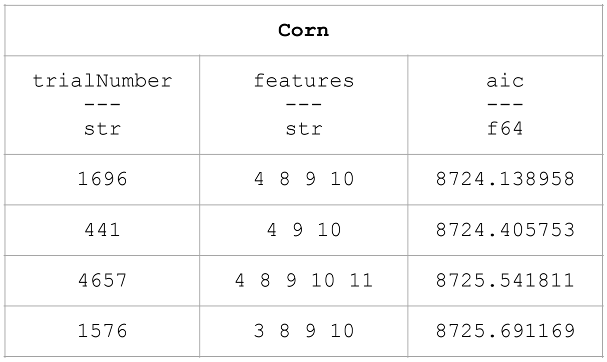
Between the commodities, the binary classification targets showed a good balance. 1 being positive return and 0 highlighting negative/flat return. I went back to calculate a proportion of Targets, which showed me that roughly 47–51% of trading days resulted in positive returns. This balance is good because it makes sure that the standard classification metrics such as accuracy maintain reliability. The near even split also suggested that commodity price movements were not systematically biased up or down over time. I would say that this was a good method for validating the suitability of logistic regression and XGBoost classifiers for supervised learning for this particular use case.

AIC-Based Feature Subset Selection

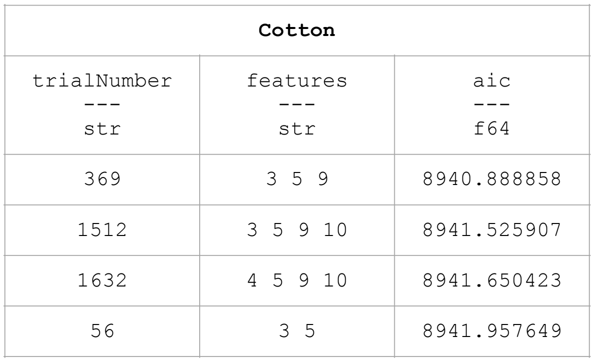
Here I have indexed the features I used for my model:



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Close Lag1	Close Lag2	Close Lag3	HML Lag1	HML Lag2	HML Lag3	OMC Lag1	OMC Lag2	OMC Lag3	Volume Lag1	Volume Lag2	Volume Lag3	Close EMA2	Close EMA4	Close EMA8



Corn		
trialNumber --- str	features --- str	aic --- f64
1696	4 8 9 10	8724.138958
441	4 9 10	8724.405753
4657	4 8 9 10 11	8725.541811
1576	3 8 9 10	8725.691169



Cotton		
trialNumber --- str	features --- str	aic --- f64
369	3 5 9	8940.888858
1512	3 5 9 10	8941.525907
1632	4 5 9 10	8941.650423
56	3 5	8941.957649

4587	4 6 8 9 10	8725.765199
1711	4 9 10 11	8725.766223
396	3 9 10	8725.825948
1697	4 8 9 11	8726.041787
1660	4 6 9 10	8726.047408
4531	4 5 8 9 10	8726.062405

Soybeans		
trialNumber --- str	features --- str	aic --- f64
922	0 9 10 13	8717.41689
923	0 9 10 14	8717.965109
505	6 9 10	8718.340386
106	9 10	8718.85203
2928	0 9 10 11 13	8718.89144
705	0 2 9 10	8719.167831
1930	9 10 12 14	8719.229885
1929	9 10 12 13	8719.254605
28	0 13	8719.278576
2594	0 3 9 10 13	8719.31603

10	9	8942.019154
414	4 5 9	8942.05857
1516	3 5 9 14	8942.090105
1449	3 4 5 9	8942.101681
1239	2 3 5 9	8942.114618
106	9 10	8942.136316

Wheat		
trialNumber --- str	features --- str	aic --- f64
922	0 9 10 13	8736.505627
923	0 9 10 14	8737.173083
2594	0 3 9 10 13	8737.582429
28	0 13	8737.616861
2889	0 7 9 10 13	8737.636924
1540	3 6 9 10	8737.668231
6256	0 3 4 9 10 13	8737.749358
2429	0 2 9 10 13	8737.749358
2714	0 4 9 10 13	8737.950651
505	6 9 10	8737.98513

Using the (AIC) / Akaike Information Criterion I performed subsets selection procedure where I reviewed the ten lowest AIC models, then selected five features for subsequent model development and examined relationships between these five selected features, along with LogReturn and CloseLag1

- **Corn**:
  - **HMLLag2** Lag-two high minus low daily prices
  - **OMCLag3** Lag-three open minus closing daily prices
  - **VolumeLag1** Lag-one daily volume
  - **VolumeLag2** Lag-two daily volume
  - **VolumeLag3** Lag-three daily volume
  - In my experience, because corn is a globally traded staples grains, it has a steady supply and demand cycle and responds to harvest conditions, global trades, and weather volitality. Therefore, the features that performed well were 3 feature lagged subsets: the closing prices for yesterday, 2 days ago, and 3 days ago (CloseLag1 to CloseLag3), the difference between daily high vs low (HML, and lagged HML), and lastly moving averages (EMA).

- **Soybeans & Wheat**:

  - **OMCLag1** Lag-one open minus closing daily prices
  - **VolumeLag1** Lag-one daily volume
  - **VolumeLag2** Lag-two daily volume
  - **CloseEMA4** Exponential moving average across four days
  - **CloseEMA8** Exponential moving average across eight days
  - These commodities included both lagged volume / price features and moving averages, which indicated higher autocorrelation in returns. According to (Cotcast.AI) Since soybeans are used as both a food and energy crop, they tend to be more sensitive to global demand shocks in energy and food consumption. Soybeans act as the prime crop for bio fuels and livestock feed globally. The repeating of features suggested to me that certain lag structures and technical indicators consistently could carry predictive information across commodity markets.

- **Cotton**:

  - **HMLLag1** Lag-one high minus low daily price
  - **HMLLag2** Lag-two high minus low daily price
  - **HMLLag3** Lag-three high minus low daily price
  - **VolumeLag1** Lag-one daily volume
  - **VolumeLag2** Lag-one daily volume
  - Cotton's optimal feature sets seemed to include much different lags compared to others. Given that cotton is a fiber, compared to grains like corn and wheat, this indicated more unique seasonal or cyclical dynamics. Cotton's demand structure, may be more closely tied to textiles and industrial production than food consumption.

Market Dynamics and Predictability

The analysis provides several insights into return dynamics:

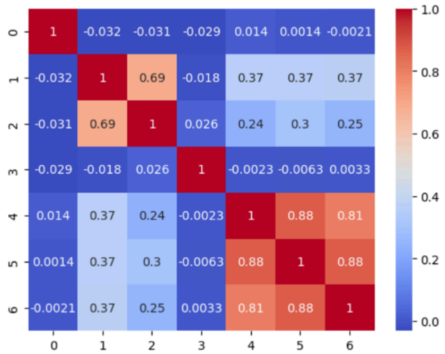
- Short-term momentum (lagged prices and EMAs) played a central role in all commodities, suggesting that recent price behavior contains modest predictive power.

- Soybeans exhibited both higher volatility and more diverse feature selection, which seemed consistent with their complex profile for demand and their exposure to risk from being used as bio fuel in addition to food.

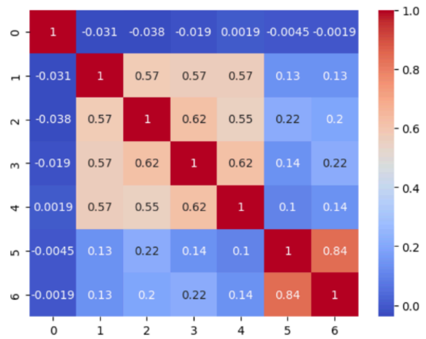
- Cotton stood apart as structurally different from the grain commodities, requiring more specialized models to capture its behavior.

Overall, while predictive power is modest, lagged and volatility-based features seemed to be consistently providing more value across the agricultural markets/commodities.

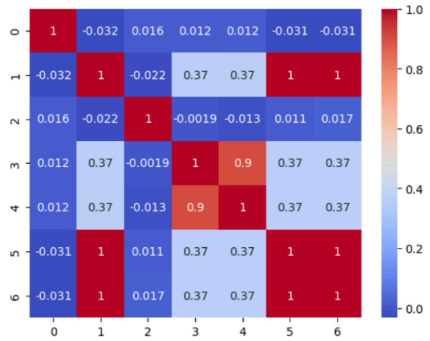
Corn



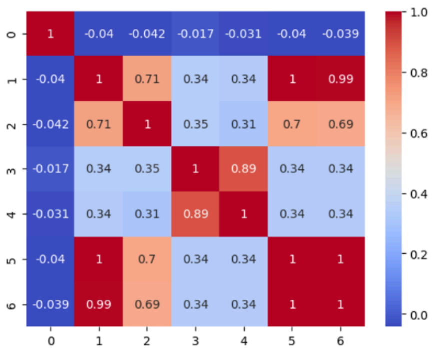
Cotton



Soybeans



Wheat



Model Evaluations

Corn

The confusion matrix had 3236 actual negative return days and 3065 actual positive return days. 2722 of the negative return days were correctly predicted, and 514 were incorrectly predicted (false positives). 2446 of the positive return days were correctly predicted, and 619 were incorrectly predicted (false negatives). So, 5168 of the 6301 days were correctly predicted, yielding an overall classification accuracy of 81.8 percent on the training data.

Corn	Predicted negative return	predicted total positive return	total totals
actual negative return	2722	514	3236
actual positive return	619	2446	3065
totals	3341	2960	6301

Cotton

The confusion matrix had 3258 actual negative return days and 3193 actual positive return days. 2782 of the negative return days were correctly predicted, and 476 were incorrectly predicted (false positives). 2729 of the positive return days were correctly predicted, and 464 were incorrectly predicted (false negatives). So, 5511 of the 6451 days were correctly predicted, yielding an overall classification accuracy of 85.4 percent on the training data.

Soybeans

The confusion matrix had 3055 actual negative return days and 3238 actual positive return days. 2407 of the negative return days were correctly predicted, and 648 were incorrectly predicted (false positives). 2718 of the positive return days were correctly predicted, and 520 were incorrectly predicted (false negatives). So, 5125 of the 6293 days were correctly predicted, yielding an overall classification accuracy of 81.4 percent on the training data.

Soybeans	Predicted	predicted total	total
	negative return	positive return	totals
actual negative return	2407	648	3055
actual positive return	520	2718	3238
totals	2927	3366	6293

Wheat

The confusion matrix had 3302 actual negative return days and 3011 actual positive return days. 2857 of the negative return days were correctly predicted, and 445 were incorrectly predicted (false positives). 2374 of the positive return days were correctly predicted, and 637 were incorrectly predicted (false negatives). So, 5231 of the 6313 days were correctly predicted, yielding an overall classification accuracy of 82.8 percent on the training data.

Wheat	Predicted	predicted total	total
	negative return	positive return	totals
actual negative return	2857	445	3302
actual positive return	637	2374	3011
totals	3494	2819	6313

References:

Ross, Sheldon M. 2011. *An Elementary Introduction to Mathematical Finance* (third edition). Cambridge, UK: Cambridge University Press. [ISBN-13: 978-0521192538]

Wilmott, Paul. 2009. *Frequently Asked Questions in Quantitative Finance* (second edition). New York: Wiley. [ISBN-13: 978-0-470-74875-6]

Cotcast AI, "Cotton, Soybeans, and Geopolitics: A Global Tug-of-War," *Cotcast AI Blog*, accessed October 3, 2025, https://cotcast.ai/blog/cotton-soybeans-geopolitics-global-tug-of-war

Miller, Tom. 2025. *451 Feature Engineering: Programming Assignment 1*. Prepared for Northwestern University, June 18.