VIP-VXU

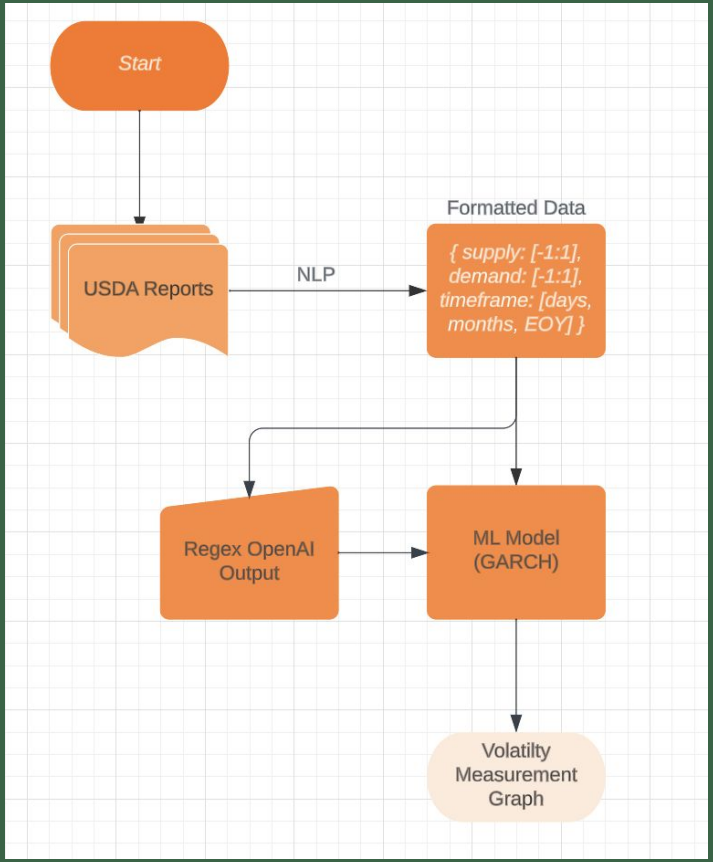
Commodity Volatility Prediction: Midterm Report

Shyam Sai Bethina, Arnav Patidar, Rohit Prasanna, Ronak Agarwal, Alex Yu-Hsin Chen, Sidharth Subbarao, Harold Huang

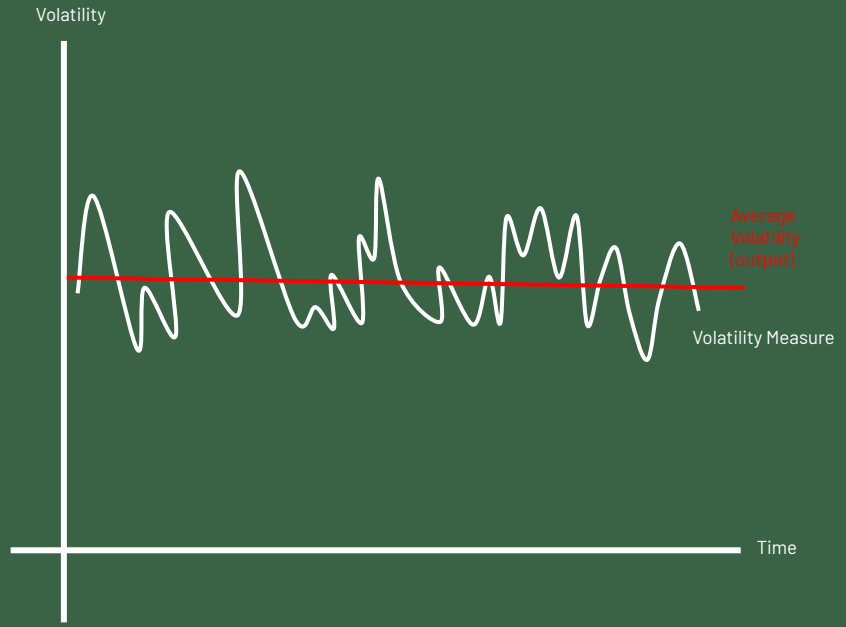
Table of Contents

General Overview

- The agriculture sector is the most critical sector for food production in the world.
- Numerous factors impact supply, demand, and ultimately the prices of the commodities that the sector pushes out. Analyzing the volatility of the most common crops is essential to understanding the commodity market.
- USDA WASDE reports are published monthly, giving namely supply & demand projections for the agricultural sector over various time periods (typically next month and next year)
- How can we utilize the language in these reports to predict the volatility of crop futures?



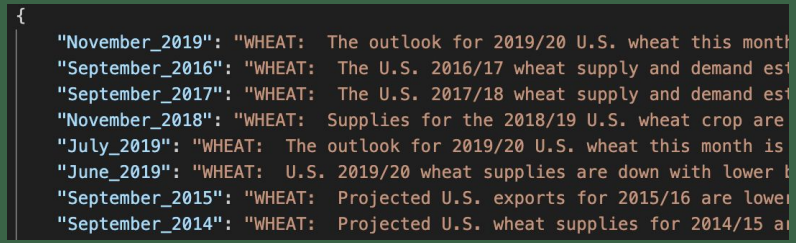
Process Flow Chart



Methodology - WASDE Reports

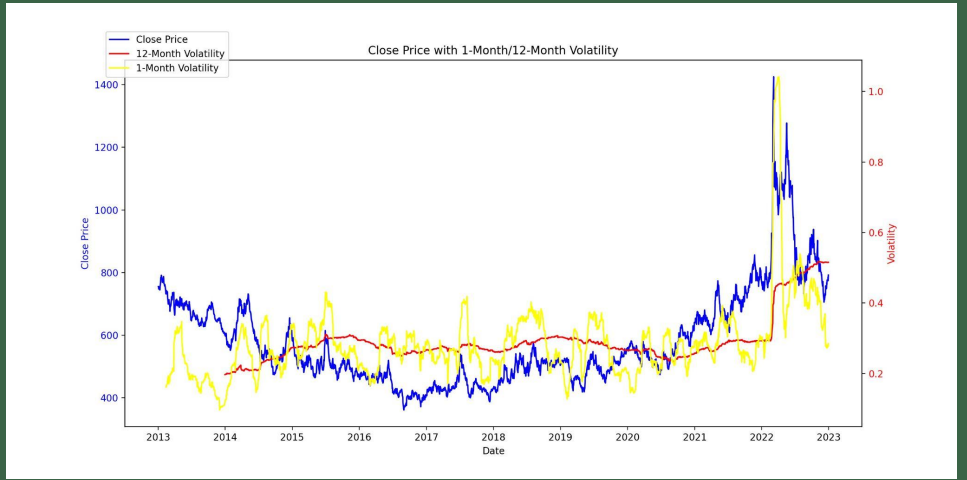
Data Collection:

- USDA World Agricultural Supply and Demand Estimates Reports (January 2013 - December 2022; 118 total)
- Used USDA ESMIS API to download reports in .XLS format
- String manipulation to get monthly highlights of soy (oilseeds), corn (coarse grains), and wheat in JSON format



Methodology - Volatility

- Scraped 1 Month, 3 Month, 6 Month, and 12 Month volatility data for wheat, corn, and soy futures
- Used Yahoo Finance to pull price data for futures from 2013-2022
- Calculated Historical Volatility based on this price data
- This is what we aim to predict through our NLP analysis



Date	1 Month Volatility	3 Month Volatility	6 Month Volatility	12 Month Volatility
2022-12-23	0.278321	0.391339	0.416158	0.515537
2022-12-27	0.275213	0.384644	0.414419	0.515372
2022-12-28	0.278197	0.382271	0.413352	0.514922
2022-12-29	0.277897	0.381248	0.412774	0.515104
2022-12-30	0.285262	0.382280	0.413637	0.515517

Methodology - LLM:

Langchain + GPT 3.5-Turbo

- Used GPT API using Langchain to analyze report highlights of wheat, soy, and corn for a conditioned response of a range from -1.0 to 1.0 for supply and demand
  - -1.0: Significant decrease within the month // 1.0: Significant increase within the month
- Using Regex expressions to format GPT output for consistency consisting of supply, demand, and time

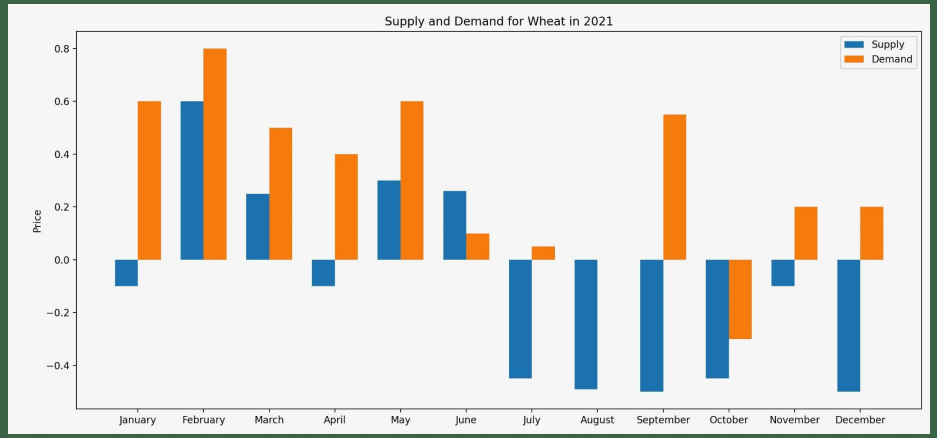
For the following USDA report: {value} that gives outlooks for wheat supply and demand, produce a decimal supply and demand rating from -1 to 1, where -1 is a strong decrease and 1 is a strong increase. Only factor outlooks that are for a timeframe of a month or less starting from this month: {date}. The report will give outlooks for multiple timeframes up to a year out. Ignore these outlooks and only consider outlooks of a month or less. Give me a response of the following format: {format}. Give me only this line after doing your analysis. Take a deep breath and approach this problem step by step.

Methodology - GARCH:

GARCH Model:

- The GARCH model estimates volatility by analyzing past price movements and squared returns, with a focus on the changing volatility in the data
- We anticipate to convert the GPT output as a return series (only numerical data) based on the appropriate time period
- It assigns different weights to past squared returns, giving more importance to recent observations, and uses these weights to forecast the conditional variance (volatility) for future periods
- By incorporating supply, demand, time period data, and sentiment analysis score as inputs, the GARCH model adapts its volatility forecasts based on both past market behavior and external sentiment information

Current Exploratory Analysis



Continued Work:

- Account for outlooks of different time frames (i.e: Longer than a month)
    - A way to to do this is getting a list of dictionaries for each WASDE Report, where each dictionary represents an outlook (ex. {'supply': -.2, 'demand': .3, 'days': EOY})
        - Needed because outlooks can be different for various time frames
- Finalize specific volatility inputs into ML model (i.e. length, timeframe)
- Work on customizing GARCH model and conducting fine tuning

Timeline

Week 9

NLP

Extract supply, demand scores using OpenAI LLM with Langchain

Week 10

Time frame Adjustment + Start ML

Accommodate for different time frame outlooks & train ML model

Week 11

Result Compilation + Model Validation

Validate methodology and model by comparing results with real world data

Timeline

Week 12

Fine-tune Model

Fine-tune model with the hopes of increasing accuracy. Potentially add more features to NLP (sentiment analysis, etc)

Week 13

Finalize Model

Continue to fine tune the model. Potentially add the given projections from USDA into ML model

Timeline

Week 14

Final Analysis

Reflect on fine-tuning changes and respond accordingly. Prepare project for final presentation.

Week 15

Presentation

Prepare final presentation to present to lab

Future Ideas

1) Expansion to diverse datasets (Reddit, FarmChat, CommodityAPI, etc.)

2) Additional features: climate, weather, geographic location,

Individual Contributions

- Shyam Sai Bethina - Worked on GPT API integration and collecting Supply/Demand scores
- Arnav Patidar - Segmented USDA Reports into respective crop, coded volatility extractor for 1, 3, 6, and 12 month periods. Langchain GPT API extraction for NLP features.
- Rohit Prasanna - USDA .XLS Commodity Report Extraction, yFinance Volatility Measurements, Consolidated project methodology
- Ronak Agarwal - Developed supply/demand visualization for past year
- Alex Yu-Hsin Chen - Background research on GARCH model, NLP methods research, calculating volatility estimates, visualization setup
- Sidharth Subbarao- Background research on GARCH model, NLP research, calculating volatility estimates
- Harold Huang - Background research on market potential of niche markets within agriculture sector, researched best futures to pull data from, developed Regex output format for ML model

Thanks!

Any Questions?