Commodities and Their Effects on the Technology Sector

Bryan Zhao, Darren Nguyen, Zayaan Atif

ECON-UB 232: Data Bootcamp

Final Project

18 December 2025

EXECUTIVE SUMMARY

Predictive Model for Vanguard Information Technology Index Fund (VGT) using commodity price data as features.

Data Sources:

- yfinance
- Reddit: 'r/news', 'r/worldnews', 'r/breakingnews', 'r/globalnews', 'r/wallstreetbets', 'r/stockmarket', 'r/stocks', 'r/trading', 'r/daytrading', 'r/economics', and 'r/economy'.

Features:

- GLD: SPDR Gold Shares (Spot Price)
- SLV: iShares Silver Trust (Spot Price)
- PPLT: abrdn Physical Platinum Shares (Spot Price)
- CPER: United States Copper Index Fund (Futures)
- USO: United States Oil Fund (Futures)
- UNG: United States Natural Gas Fund (Futures)
- WEAT: Teucrium Wheat Fund (Futures)
- SOYB: Teucrium Soybean Fund (Futures)
- CORN: Teucrium Corn Fund (Futures)

Models:

- Logistic Regression (LR)
- K-Nearest Neighbors (KNN)
- Extreme Gradient Boosted Classifier (XGB)
- Artificial Neural Network (ANN)
- Sentiment Analyzer “tabularisai/multilingual-sentiment-analysis”

The best-performing model was the Extreme Gradient Boosted Classification. It had a precision score of 65.8% and an F-Beta score of 68.5%.

The most important features for VGT are precious metals such as copper, silver, and gold.

NTRODUCTION

Tools for investment analysis are ever-changing. Investors are finding new ways to gather, analyze, and predict movements within the stock market. Within the discipline of quantitative finance, machine learning and neural networks have emerged as crucial methods for data analysis, attracting significantly more attention in recent years.

One overlooked factor when investing in stocks is commodity prices, some of which play an important role in the technology sector. For example, gold, silver, copper, and platinum are all utilized to manufacture technological infrastructure. As a result, a trading strategy focused on the technological sector could be developed based on these commodities.

The goal of this project was to predict movements in the technology sector using information about commodity prices.

Essential Questions

1. Which model is most effective in predicting an index's movement based on commodity price changes?
2. Which commodities are most important to the technology sector?
3. What is the most important commodity to market sentiment, and how does this relate to our models' findings?
4. Can we develop a trading strategy using our model?

DATA DESCRIPTION

Commodities

Because we were analyzing the movement of different commodities, we used ETFs for price data. These values represented either the spot price (how much it costs to buy the commodity now) or the futures price (how much it will cost to buy the commodity in the future). We used the following 9 ETFs, spanning across precious metals, fuels, and agriculture.

- GLD: SPDR Gold Shares (Spot Price)
- SLV: iShares Silver Trust (Spot Price)
- PPLT: abrdn Physical Platinum Shares (Spot Price)
- CPER: United States Copper Index Fund (Futures)
- USO: United States Oil Fund (Futures)
- UNG: United States Natural Gas Fund (Futures)
- WEAT: Teucrium Wheat Fund (Futures)
- SOYB: Teucrium Soybean Fund (Futures)
- CORN: Teucrium Corn Fund (Futures)

We used the Vanguard's Information Technology Index Fund (VGT) as a proxy for the technology sector.

YFinance

Using the YFinance API, we extracted the ten-year historical price data on Vanguard Information Technology (VGT) and our nine commodities of interest. With the price data, we calculated the percent changes to reflect relative movement for each ETF. We applied a Standard Scalar to utilize in some of our models. Then, because our goal was to predict VGT's direction of movement, we represented any positive percent change as a 1 and represented any negative (or zero) percent change as 0. This was our target column. Finally, we applied a train, test, and split to the data to accurately develop our models.

Reddit

Using the Praw API, we scraped the top 500 posts of the past month on the subreddits 'r/news', 'r/worldnews', 'r/breakingnews', 'r/globalnews', 'r/wallstreetbets', 'r/stockmarket', 'r/stocks', 'r/trading', 'r/daytrading', 'r/economics', and 'r/economy'. It was necessary to use this many subreddits to get enough data to extract any valuable information from. We returned only posts that mentioned commodities, as well as their titles and bodies, to run sentiment analysis on.

MODELS & METHODS

Supervised Learning Classification Models

We used three different supervised learning classification models: Logistic Regression (LR), K-Nearest Neighbors (KNN), and Extreme Gradient Boosted Classifier (XGB). Using SciKit Learn, we created pipelines for each of the three models, fitted them to the train data, and scored them to determine which model performed best. To determine the optimal parameters, we employed GridSearch on the nearest neighbors for KNN and max depth for XGB.

Neural Network for Classification

In addition to our three supervised models, we created an artificial neural network (ANN) using PyTorch. In our Sequential object, we implemented a three-layer model with 128 neurons, 64 neurons, and 32 neurons, respectively. Passing our nine input features through, we had one output feature.

Evaluating Models

To evaluate our models, we used both a precision and an F-Beta score. This study only applies to buying VGT shares, so it would be bad to predict VGT increases when it actually goes down (incurring losses), but not as bad to predict VGT decreases when it actually goes up (missing gains). Because investors want to avoid losses, we chose precision as a scoring metric. However, if investors avoid all losses, they would never make any money. The F-Beta score provided a metric that mixed both precision and recall, balancing loss aversion and gain aversion. We used a beta of 0.5 to prioritize precision.

Determining Feature Importance

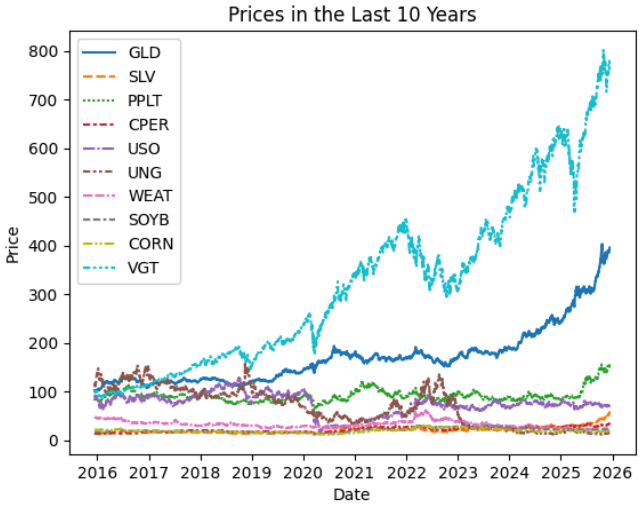
To compare commodities and determine which were useful features, we used permutation importance. This method shuffles feature values around and calculates the effects on model performance. The most important features are the ones that drop performance the most when randomized.

Neural Network for Sentiment Analysis

We used the “tabularisai/multilingual-sentiment-analysis” to analyze the sentiment of the title and body of each Reddit post. We used an existing model from HuggingFace as it was already trained on a lot of data, and this would be a lot more effective than trying to create our own. The model ranked the texts as either “Very Negative”, “Negative”, “Neutral”, “Positive”, or “Very Positive”.

RESULTS & INTERPRETATIONS

We plotted closing price data from the last 10 years for VGT and the 9 commodities.



Baseline

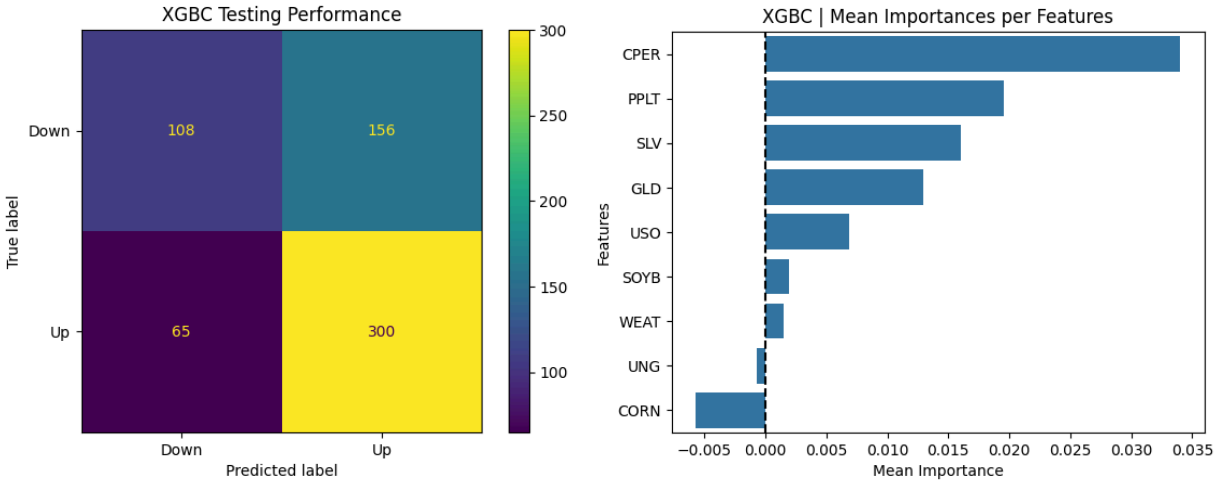
VGT had 1436 positive days out of 2514 trading days. The baseline for our models was 57.1%.

Extreme Gradient Boosted Classifier (XGBC) | (Optimal Max Depth: 1)

Precision Score: 65.8% | F-Beta Score: 68.5%

High Mean Importances: Copper, Platinum, Silver, Gold

Low Mean Importances: Corn, Natural Gas, Wheat, Soybean, Oil

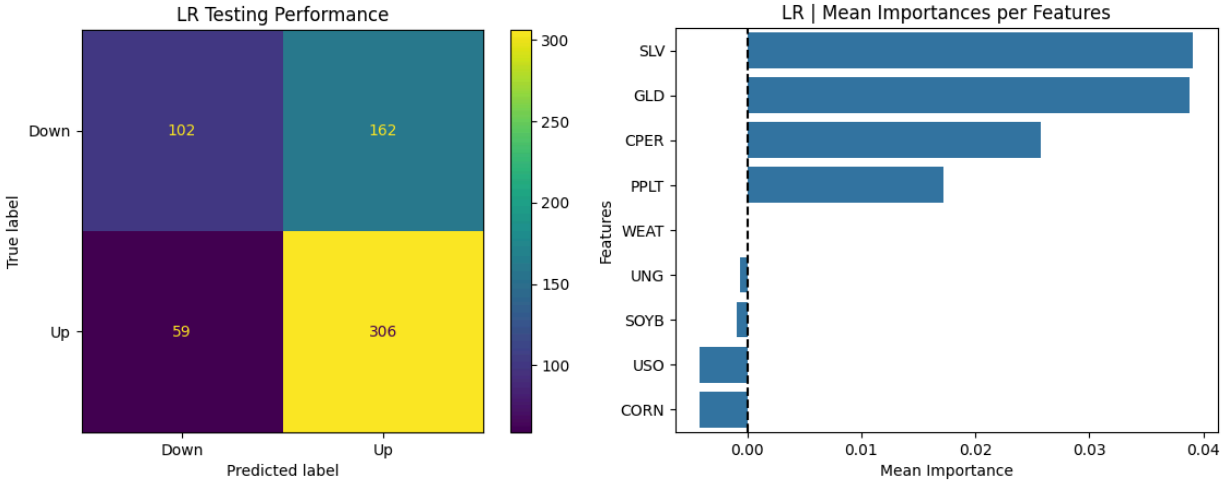


Logistic Regression (LR)

Precision Score: 65.4% | F-Beta Score: 68.4%

High Mean Importances: Silver, Gold, Copper

Low Mean Importances: Corn, Oil, Soybean, Natural Gas, Wheat

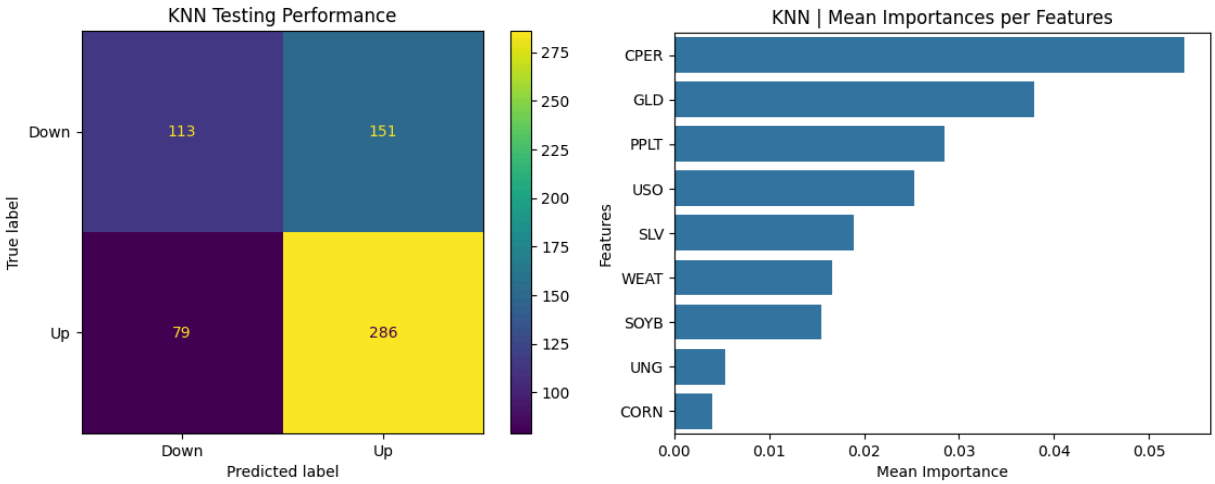


K-Nearest Neighbors Classification (KNN) | (Optimal N-Nearest Neighbors: 9)

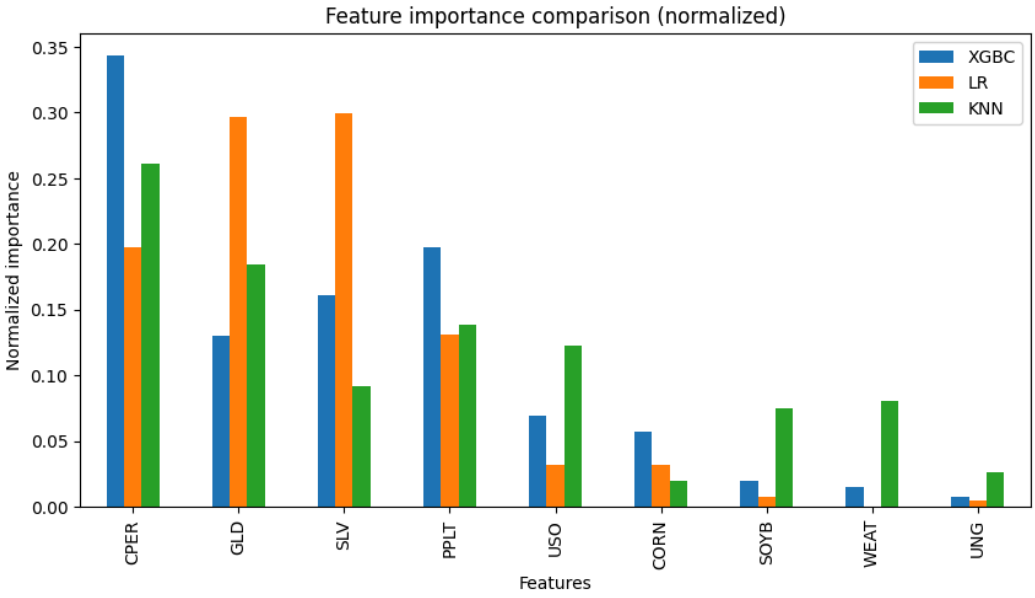
Precision Score: 65.4% | F-Beta Score: 67.7%

High Mean Importances: Copper, Gold, Platinum, Oil

Low Mean Importances: Corn, Natural Gas, Soybean, Wheat, Silver

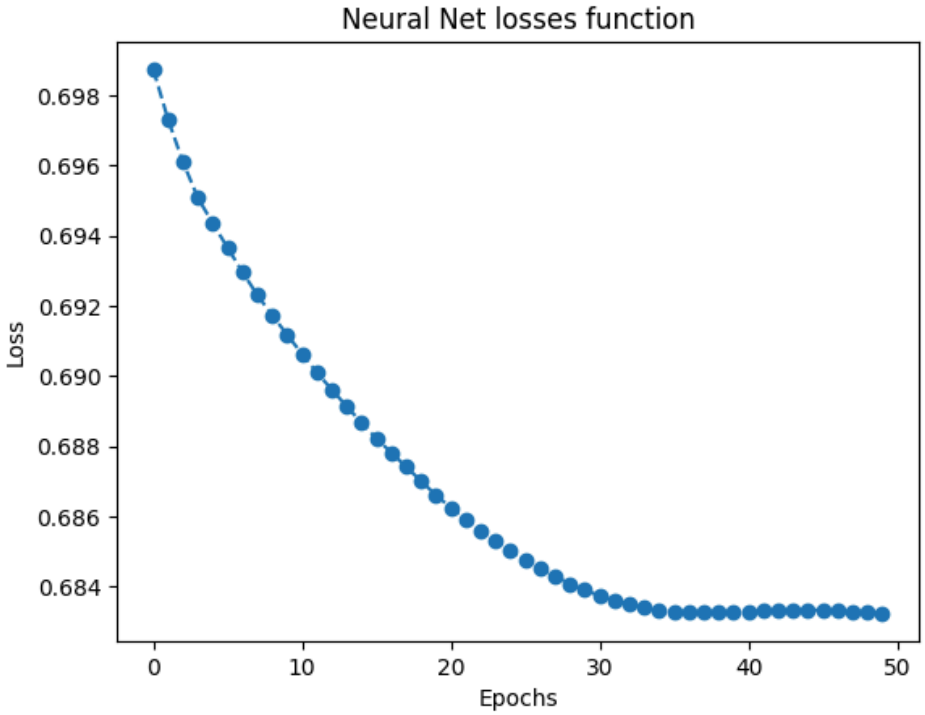


Comparing Feature Importances

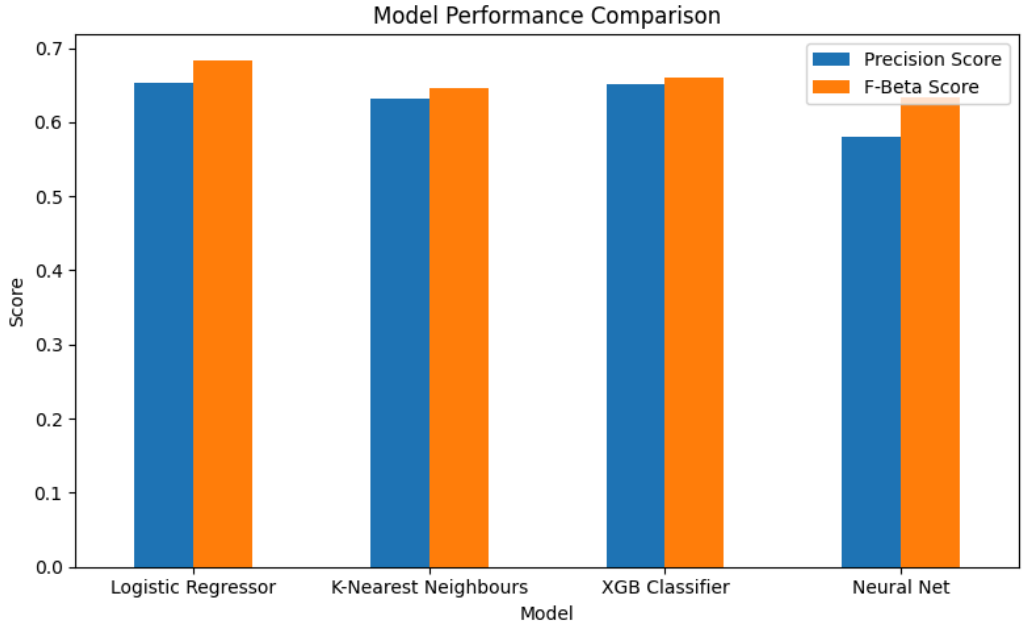


Artificial Neural Network (ANN) | (3 hidden layers, ReLU activation functions)

Precision Score: 58.0% | F-Beta Score: 63.3%

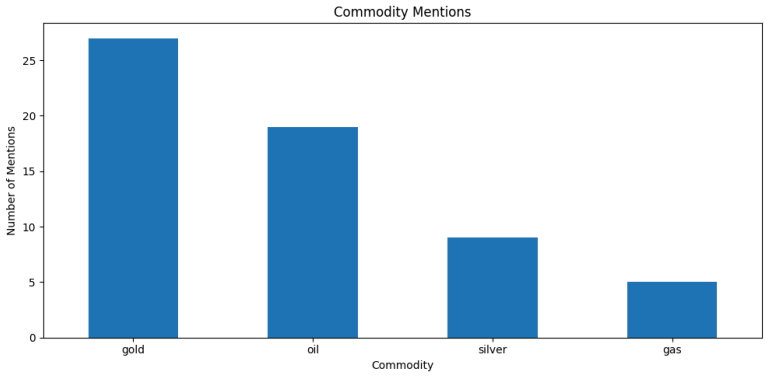


Comparing Performance Scores

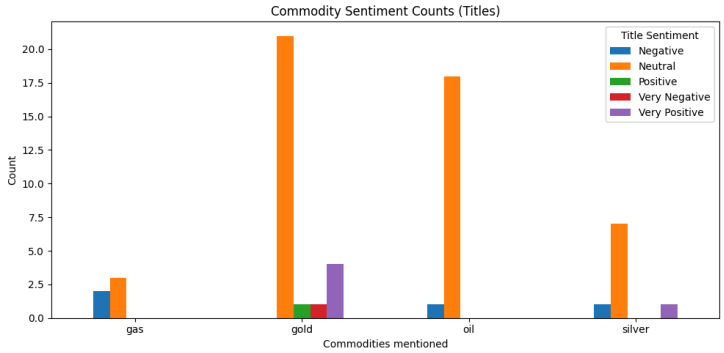


Sentiment Analysis

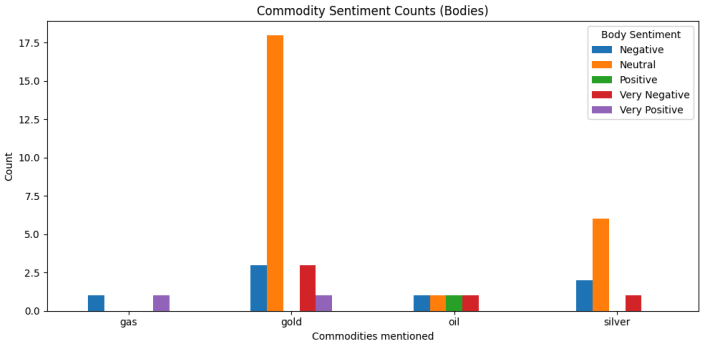
Gold was the most mentioned commodity, followed by oil and silver. Surprisingly, there was a lot of discussion about gold and silver, but not as much about copper, which was the most important variable for most of our models. However, gold and silver may be highly mentioned due to idiomatic phrases or other uses, not just stock-related.



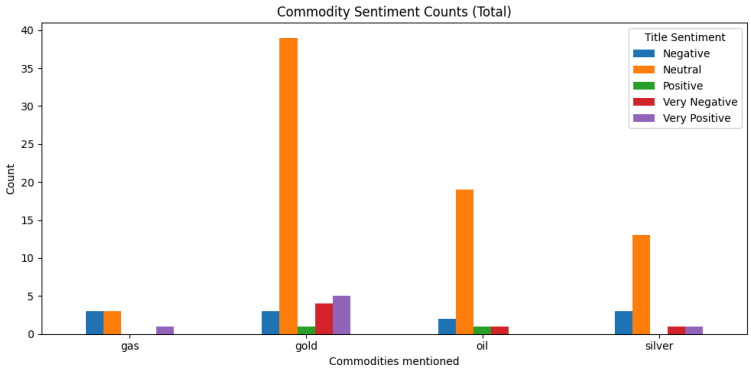
Titles are generally more neutral, and the model reflects this with the overwhelming majority of predictions being neutral.



Bodies can be more complex in their sentiment, shown by the wider range of sentiments. However, the analyzer was not very good at determining context, and so there are still many neutral predictions.



The combined sentiment of all titles and bodies is displayed below.



CONCLUSION & NEXT STEPS

Which model is most effective in predicting an index's movement based on commodity price changes?

The best-performing model was the Extreme Gradient Boosted Classification. It had a precision score of 65.8% and an F-Beta score of 68.5%. However, all 3 supervised learning models had very similar scores, varying by only a few tenths of a percentage point. Meanwhile, the artificial neural network for classification performed the worst.

Which commodities are most important to the technology sector?

Across all supervised learning models, precious metal commodities typically had the largest mean importance, with copper, silver, and gold being the most important. The least important features were the agricultural commodities, as we expected. Interestingly, fuel commodities such as natural gas and oil did not seem very important, contrary to our initial beliefs.

What is the most important commodity to market sentiment, and how does this relate to our models' findings?

According to our webscraping and sentiment analyzer, gold, oil, silver, and gas were the most mentioned commodities. Interestingly, copper, the most important feature in our model, was not a commonly mentioned stock.

Can we develop a trading strategy using our models?

Although our models performed better than the baseline, about 65% precision compared to 57%, our models would not directly be used for a trading strategy. It is important to note that we used the closing prices for VGT for the same days as our commodity prices. This makes our models more focused on predicting the movement of the technology sector given the commodity price changes for that day, rather than directly predicting future VGT prices. However, the results showed important precious metals to track, something that can be further explored in a trading strategy.

Conclusion

Through optimizing and evaluating multiple models to predict the movement of the technology sector based on the movement of commodity prices, we found that the Extreme Gradient Boosting Classifier was the most effective model in predicting the movement of the technology sector, whilst minimizing type I errors. Additionally, in terms of feature importances, precious metal commodities, such as copper and silver, were the most important, somewhat similar to what the sentiment analyzer found.

Next Steps

To further improve our model, we want to expand our data scope and sentiment analyzer. First, our data was very narrow in scope and lacked a prospective view on the stock. Implementing more data, such as discounted cash flows, profitability, and revenue lines, would improve the accuracy of our model. Additionally, our sentiment analyzer was not fine-tuned for our goal in analyzing stocks and commodities specifically, making our market research difficult and inaccurate in certain scenarios. Developing a financial-tailored sentiment analyzer would allow us to more accurately scrape news sources to gather prospective data on VGT, other stocks, and ETFs.

In minimizing errors and selecting a more effective model, we believe that we can utilize these findings in developing a trading strategy.

** Note: Figures in the code may be different from the paper and presentation due to new market data. The data shown in the pictures of this paper and the presentation came from December 10th, 2025.