CAPSTONE SPRINT 2

ANOMALY DETECTION OF CONGRESS TRADES

Tiffany Chu

Problem Area and Opportunity

Insider Trading

- US Congress members create policies and can leverage non-public information for financial gain,
- The STOCK act requires them to report their transactions

Opportunity

- There is a need to identify deviations that could suggest insider trading activities and stock manipulation

Parties Impacted:

Retail Investors

Regulatory Bodies

Financial Analysts/ investors

Preprocessing and EDA

1

Merging Data

Matching S&P data with the transaction dates of congress data

2

Scaling and one hot encoding

Scaling using robustscaler so its not dominated by outliers

Using one hot encoding to turn categorical data to numerical

3

Time Series analyses

Plotting market performance data (volume and closing price - adjusted)

Autoencoding

Anomaly Detection using Unsupervised ML Algos:

Hierarchical Clustering

- Find patterns in unlabelled data and groups them into clusters

Isolation Forest

- Recursively splits data into subsets, anomalies can be isolated with fewer partitions

Autoencoders

Data to be Analyzed

1. Congress Trades

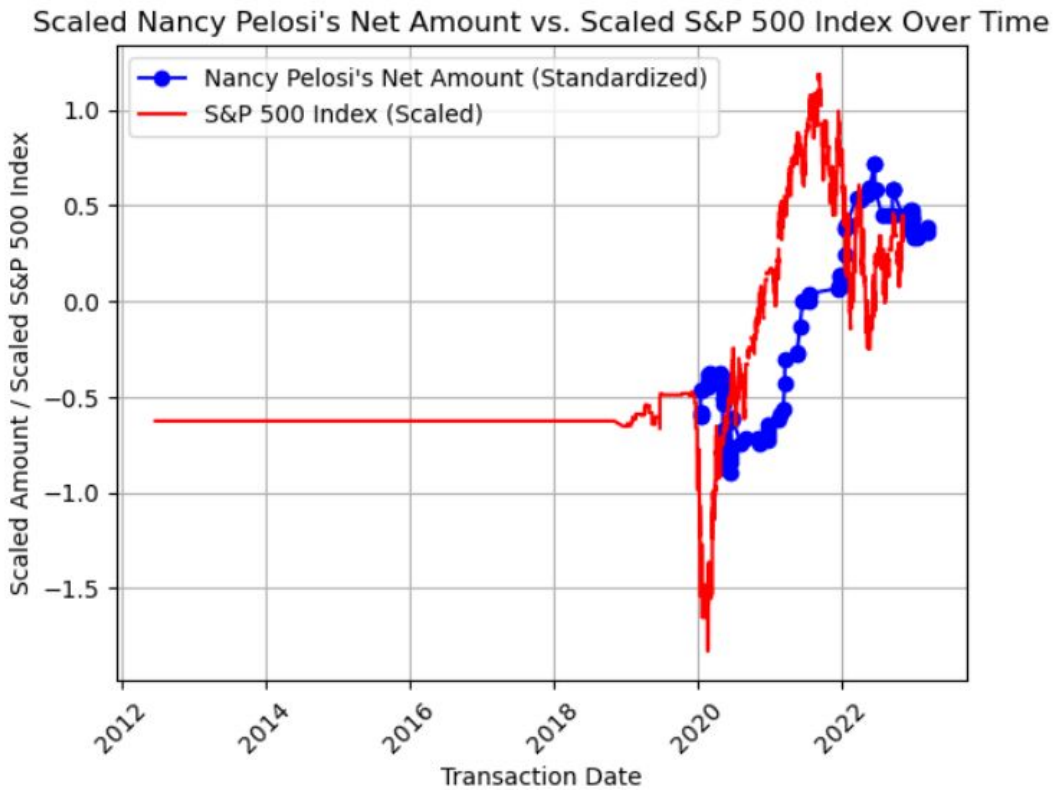
- 01 — Transaction Type (buy, sell) and gains
- 02 — Transaction Date mm/dd/year
- 03 — Industry and Sector
- 04 — Stock owner and Congress name

2. S&P Data

- 01 — Event date time frame
- 02 — Industry and companies affected

Feature Eng: Net invested, days before declaration, upper bound invested

Time Series findings



Nancy Pelosi Net investments vs Market

An example here is:

Top earning congress members **know to sell when the market is doing well, and invest when its weaker, meaning prices are lower.**

This makes them appear to predict the timing of the market

Next Steps

Research Results

Use domain knowledge to interpret the anomalies. **confirm through manually searching anomalies.**

- Also: accuracy, recall, precision, ROC curve, AUC PR (area under) are good metrics for investigation

Visualize

Visualize the data surrounding the anomalies to identify any patterns or trends that might explain them

Iterative Improvement

Improve the anomaly detection system's performance and effectiveness by **repeatedly refining hyperparameters**, which might be a challenge since there's no ground truth for validation.

Thank you!

Github:
https://github.com/tiffchu

Stock News Dataset:
https://www.kaggle.com/datasets/parsabg/stocknewseventssentiment-snes-10

Congress Trades Dataset:
https://senatestockwatcher.com/api