Enhanced Stock Prediction and Algorithmic Trading System Using Cointegration Testing and Stock Chains

-Ryan Mennemeier

-September 2024

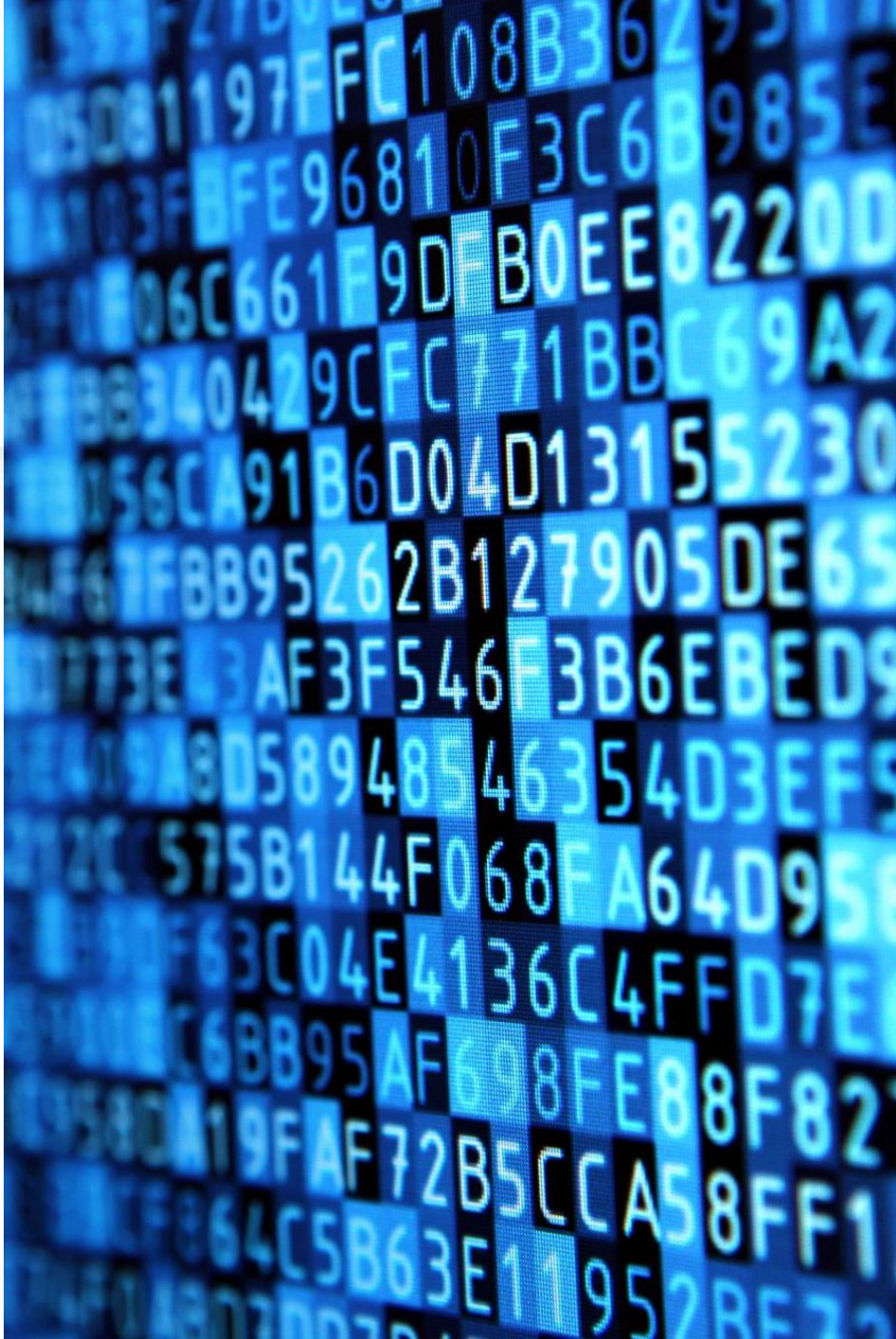
Project Overview

- This project involved taking historical data for 10 “core” stocks, analyzing them in detail, performing baseline modeling for each, then doing cointegration testing and construction of stock chains with a set of generated secondary stock data as well as exogenous data we had gathered.

Data Collection

Using yfinance:

- Collected 5 years' historical data for each of our core stock data.
- Generated a list of 200 "secondary" stocks also with the 5 year history to be used in the cointegration step.
- Acquired exogenous data.



Data Preprocessing

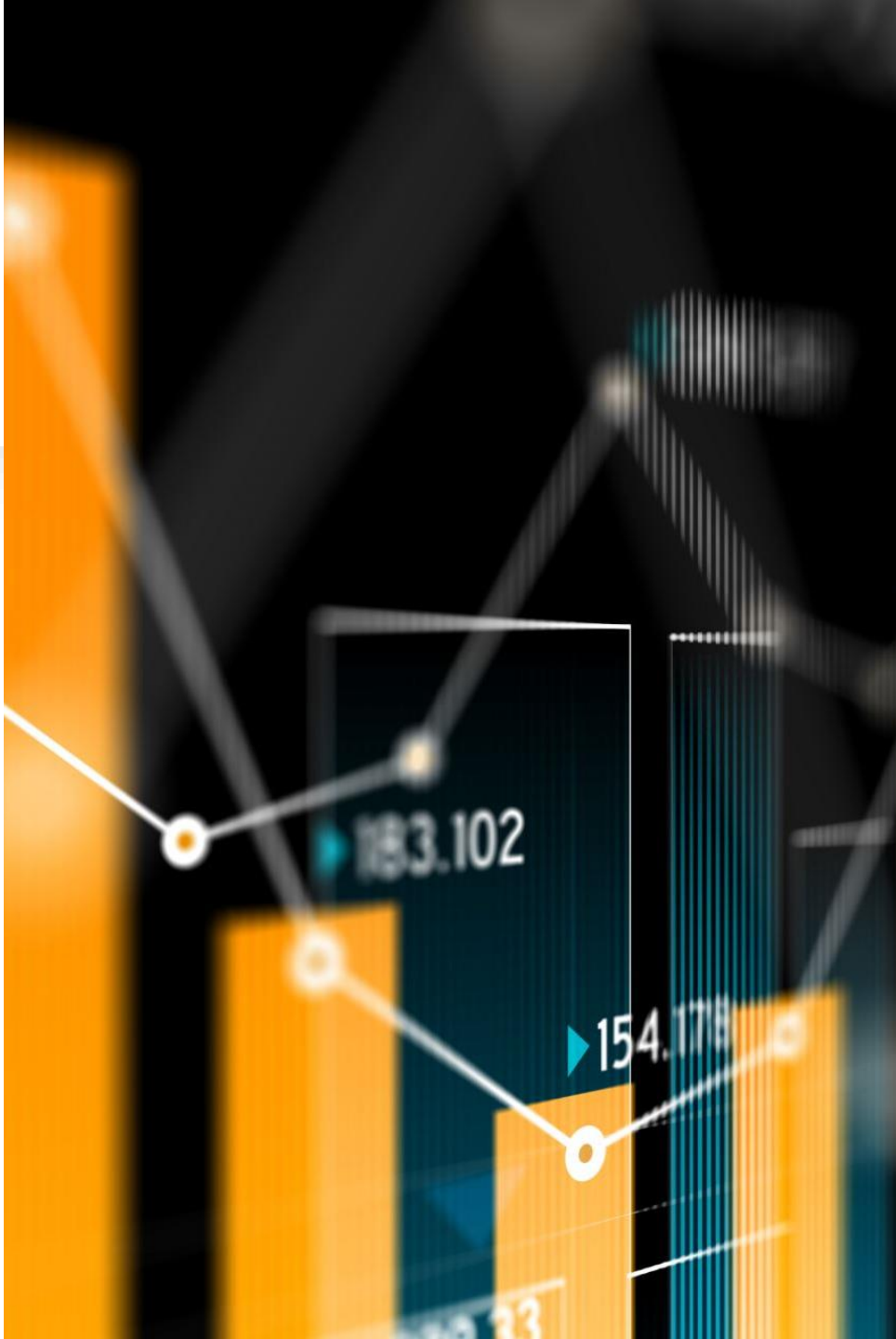
- Cleaned the data, handled missing values, scaled features, and ensured proper formatting.
  - Went through each dataset, and ensured missing values were imputed and none were removed.
  -Preprocessed using zscore and log transform,
  scaled at the end.

Feature Engineering

- Generated technical indicators like SMA, EMA, MACD, and lag features for analysis.
  - Researched value-add indicators and input them in both the core and secondary stock dataframes.
  - Used a 3-day lag window, to be able to show a small forecast window at the end for the stock chains.

EDA and Baseline Model

- Performed exploratory data analysis (EDA) for each core stock. I looked for any discernible trends.
- For the baseline model I landed with Linear Regression to be able to see how the core stocks were predicting based on my preparation work.



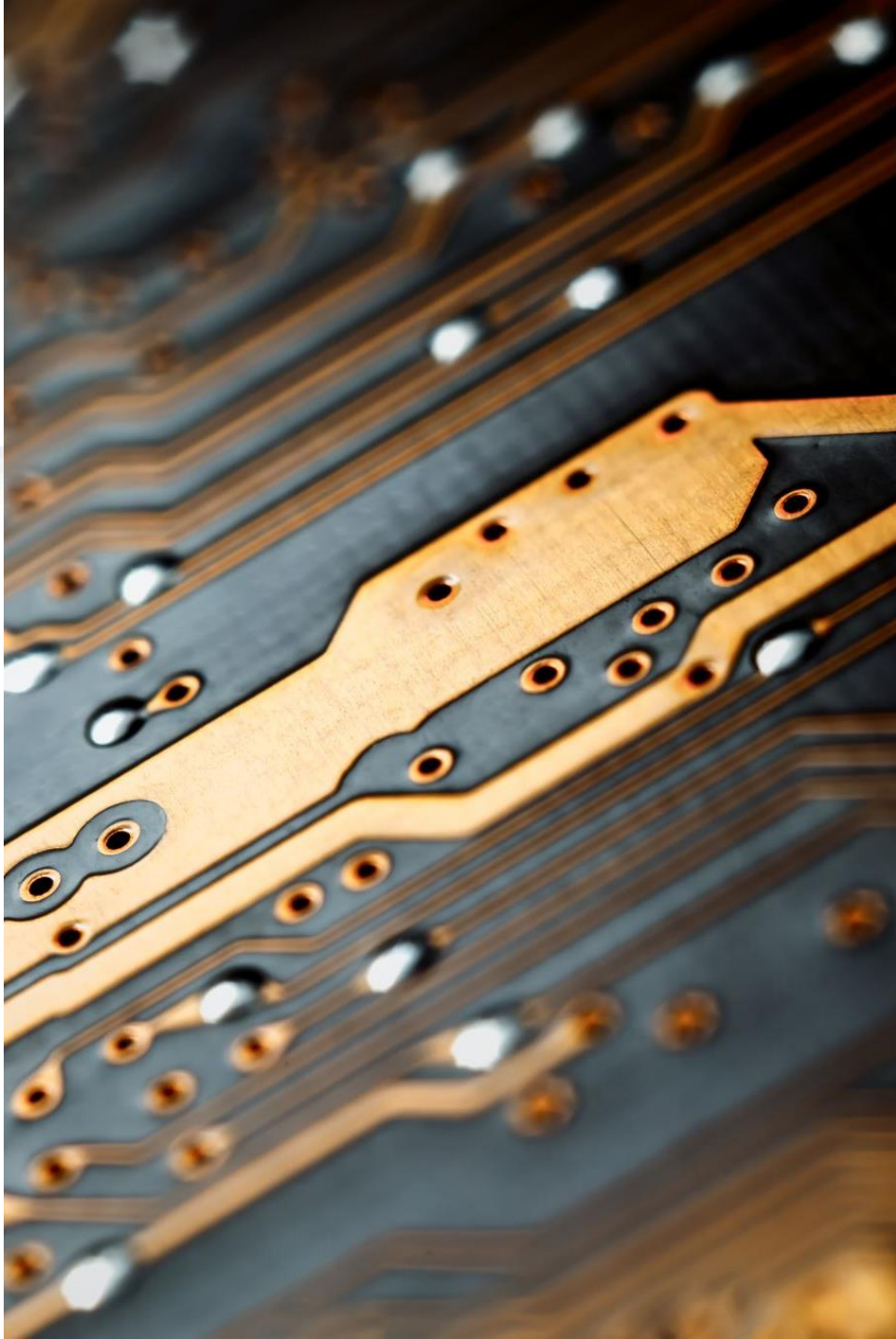
Modeling Approach

- Initially used Lasso regression with the full feature set (153), to very mixed results. Performed multiple renditions until I landed on Ridge for the duration.

Feature Optimization

- This part was really important, as we were able to narrow down our list of features from 153 to an optimized list of 25!

- PCA
- RFE
- VIF
- MI Scores



Evaluation Metrics

- Tracked RMSE, MAPE, R2, and adjusted R2 scores to assess model performance for both training and testing sets for the Ridge model.

Ridge Model Performance

-See table for full range of metrics for the core stocks on the Ridge model.

Ticker	Train RMSE	Test RMSE	Train MAPE	Test MAPE	Train R2	Test R2	Train Adj R2	Test Adj R2
AAPL	0.0006	0.0007	0.6 %	1.0 %	0.999	0.999	0.999	0.999
AMZN	0.001	0.001	2.1 %	0.3 %	0.999	0.999	0.999	0.999
GOOG	0.074	0.09	68 %	40 %	0.95	0.50	0.95	0.50
MA	0.0007	0.0008	0.05 %	0.05 %	0.999	0.999	0.999	0.999
META	0.0017	0.0018	0.6 %	0.1 %	0.999	0.999	0.999	0.999
MSFT	0.0005	0.0008	0.15 %	0.05 %	0.999	0.999	0.999	0.999
NVDA	0.002	0.007	0.08 %	8.2 %	0.999	0.999	0.999	0.999
PFE	0.0005	0.0005	0.04 %	0.03 %	0.999	0.999	0.999	0.999
PG	0.0004	0.0004	0.7 %	0.1 %	0.999	0.999	0.999	0.999
TSLA	0.01	0.01	1.6 %	1.5 %	0.999	0.996	0.999	0.996

Improvement and Iteration

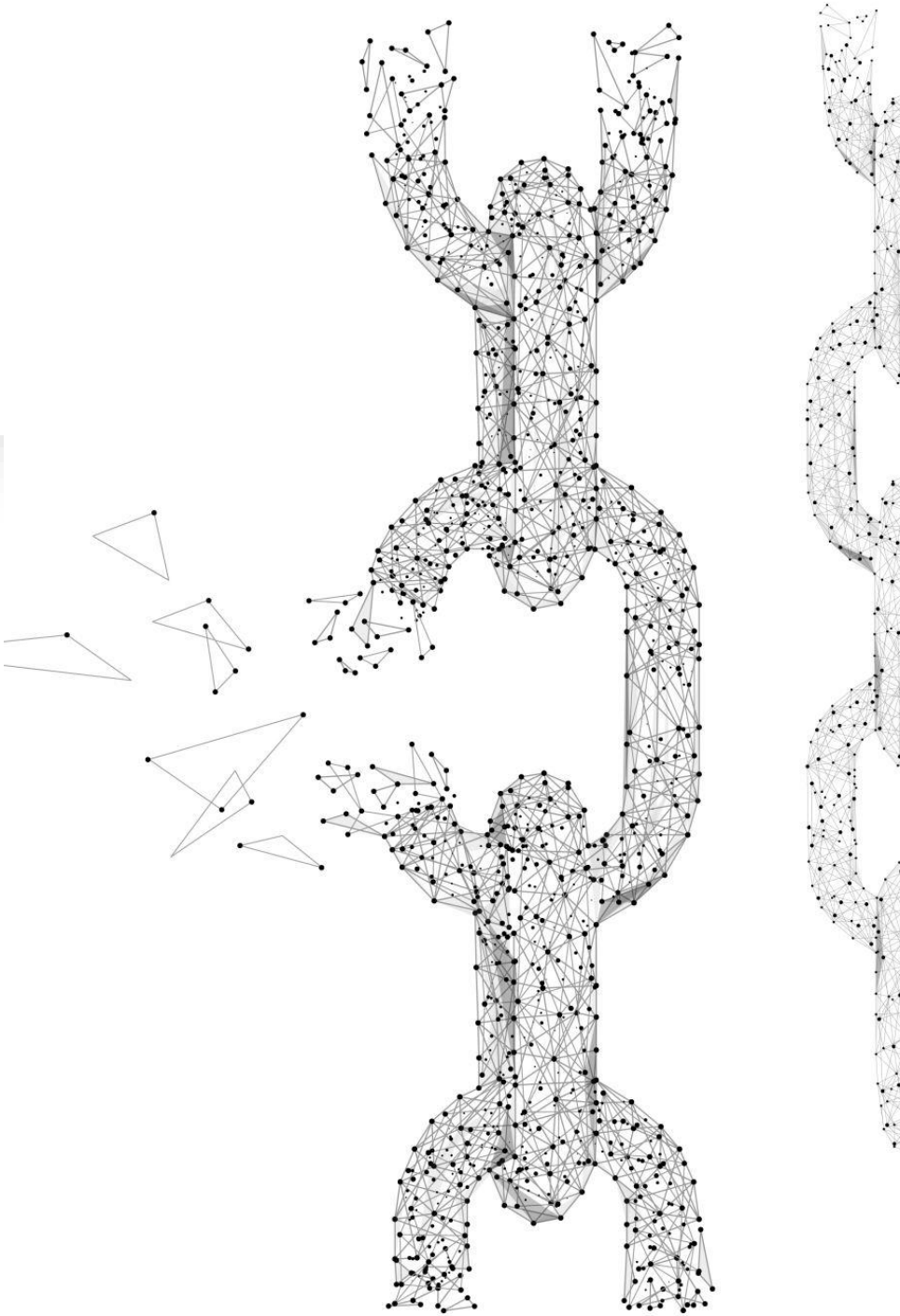
Cointegration Testing

- Tested cointegration between the core stocks and secondary stocks to find pairs, as well as the exogenous chosen features.
- Table at right are the pairs chosen after Layer 1 cointegration.

Stock Ticker	Sec Stock Pairs	Exo Feature Pairs
AAPL	3	0
AMZN	1	0
GOOG	6	3
MA	40	5
META	0	0
MSFT	8	0
NVDA	4	0
PFE	0	1
PG	33	4
TSLA	5	0

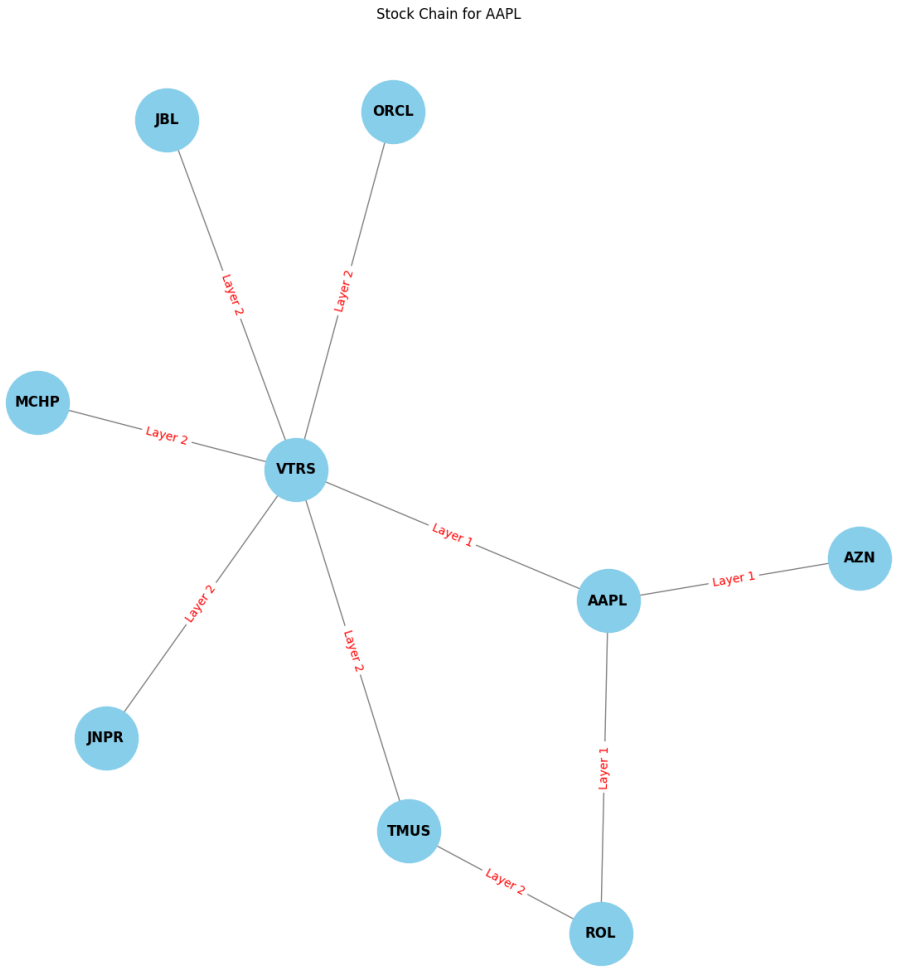
Stock Chains

- Constructed stock chains (layered cointegration testing) from the pairs gathered from the first testing performed.



Final Stock Chains

-Completed AAPL Stock Chain after Layer 4.



Future Steps

- **Potential future steps include:**

  -Developing a robust Trading Strategy for each Stock Chain, then backtesting for accuracy.

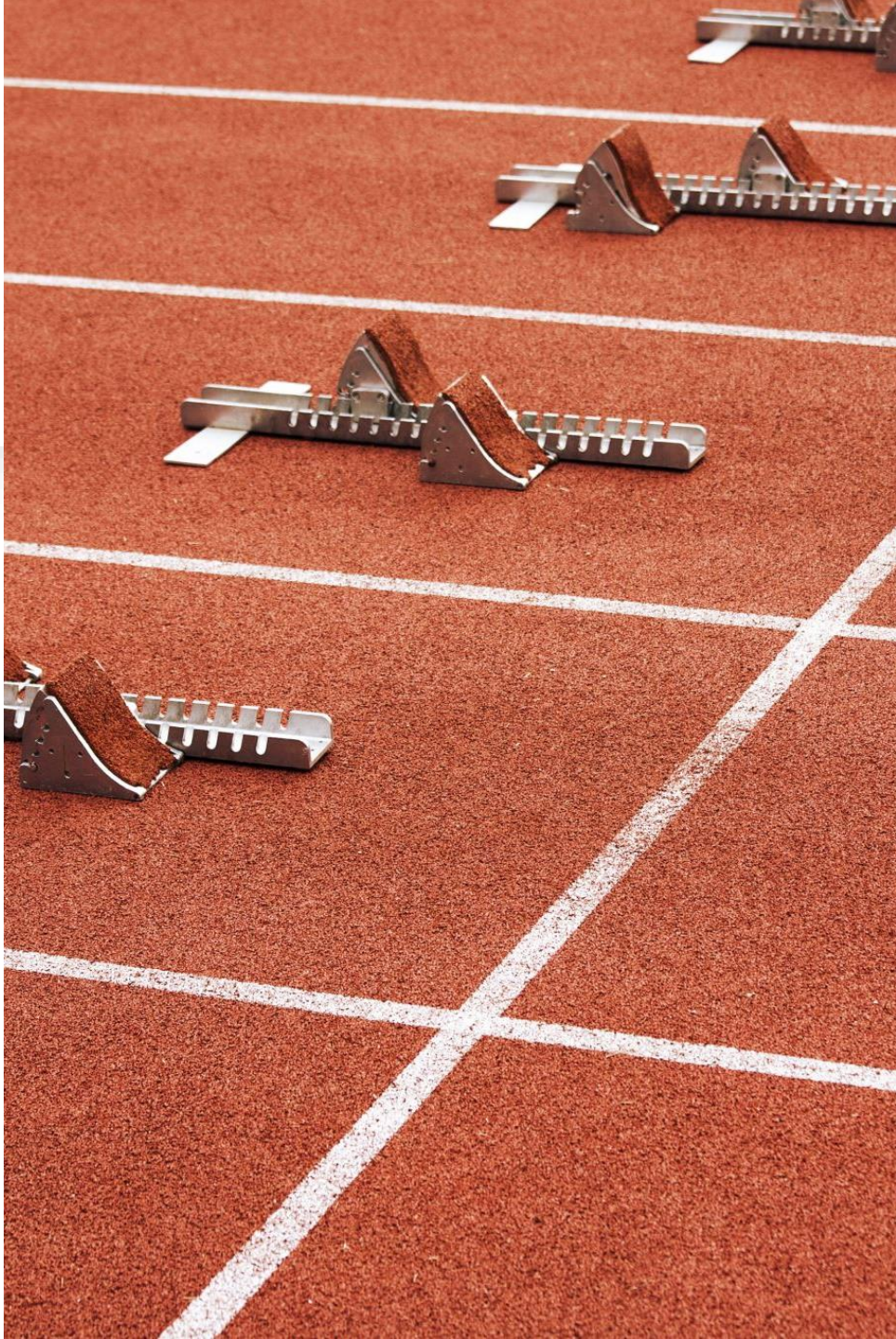
  -Run the stock chains through GRU for forecasting.

  -Once backtested deploy through the stock chains through a simulator for a set amount of time and optimize the strategies.

Challenges

There were several hurdles in this project:

- Foreign topics.
- Time was not my friend.
- The unforeseen bug or feature snag, causing longer than intended delays in progress.
- Project scope creep.



Results

Achieved overall great metrics for our baseline model.

Acquired a good optimized feature list to use going forward.

Have found that most of our core stocks are capable of cointegration, and have built full stock chains through the 4th Layer. Am confident for the next steps.

Recommendations

- The methods performed shows that the process works as intended, just needs more time.
- Acquire a better API (subscription) for finance data with more indicators and history to embolden the existing data, and replace the core stocks that didn’t perform well in cointegration.
- Be patient, the real payoff will be when the stock chains have their strategy attached and are performing in the simulator.

Thank you Mr. AJ, I appreciate all of your time!

Thank You!
Any Questions Today?