Citadel Spinoffs

Final Presentation for: "*Data Driven Methods in Finance, IEOR 4576*"

Presented by: Minas Emiris, Loris Domeneghetti, Seth Younger, Martin Roux, Lenny Schwarz

Outline of Presentation

1. Introduction to Datasets
2. Data Engineering
3. Feature Ideation & Engineering
    a. Technical Factors
        i. Bollinger Bands
        ii. SARIMAX
    b. Alternative Factors
        i. Analyst Recommendations
        ii. Starmine Models
4. Combination of Models (Black Litterman Model)
5. Final Disclaimers

Introduction to Datasets

Overview of Modelled Data:

1. *Stocks (50 assets)*
   a. ***Fundamental:*** *Valuation, Size, Operating Efficiency, Profitability, Risk, Corporate Activity*
   b. ***Technicals:*** *Liquidity, Price-Based, Overall Market*
   c. ***Alternative Factors:*** *Ratings, Revisions, Distributions*
   d. ***Economic Factors:*** *Inflation, GDP Growth*
   e. ***Alternative Social Responsibility: ESG Score***
2. *ETFs (50 assets)*
   a. ***Fundamentals: Scope, Valuation, Size***
   b. ***Technical: Liquidity, Price-Based, Overall Market***
   c. ***Alternative: Rating, Revisions, Distributions***
   d. ***Economic Factors: Inflation, GDP Growth***
   e. ***Social Responsibility: ESG***
3. *Crypto (10 assets)*
   a. ***Fundamentals:*** *Scope, Market Size*
   b. ***Technicals:*** *Liquidity, Price-Based*
   c. ***Alternative:*** *Scope, Volume of Mining*

General Methodology:

1. ***Preliminary Idea Creation:*** *Academic Papers, Bibliography*
2. ***Feasibility Study:*** *Check if forming the factors was achievable*
3. ***Feature Creation & Testing:*** *Data Sourcing, Engineering, Testing*
4. ***Assumptions Study:*** *Replicability Studies of factors*

Data Provider	Advantages	Disadvantages
yahoo! finance	• Free to use • Well-documented • Straightforward API	• Limited advanced analytical tools • Data delays for real-time information
REFINITIV	• Comprehensive and high-quality data • Global market coverage • Advanced analytical tools	• High cost • Complexity for new users • Limited crypto data compared to stocks and ETFs
FACTSET	• Deep and extensive datasets • Advanced portfolio and risk analytics • Strong integration capabilities	• High subscription cost • Steeper learning curve • Primarily tailored for professional investors and institutions

Data Leveraged:

1. ***Yahoo Finance: Price Signals were used in conjunction to other platforms***
2. ***Refinitiv: Fundamentals, Technicals, Alternative Factors***
3. ***FactSet: Sentiment Scoring, Barra Regression***
4. ***MarketWatch: Analyst Ratings, Revisions, Distributions***

Data Engineering

General Methodology

1. **Exporting from relevant datasets using API & SQL queries (Factset, Refinitiv, Yahoo! Finance, MarketWatch)**

1. **Distribution Plot Creation (Data Integrity, Comparative Analysis, Histograms & Box Plots, Time-Series Analysis)**

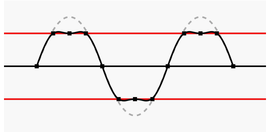
1. **Data Cleaning (Stage 1)**
   a. Remove non-trading days
   b. Remove non-trading history

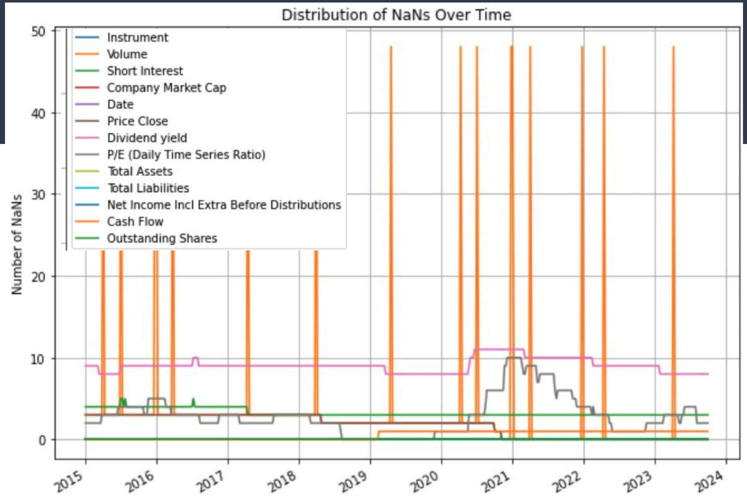
```
Apple_SPY_stock_price.tail(10)
```

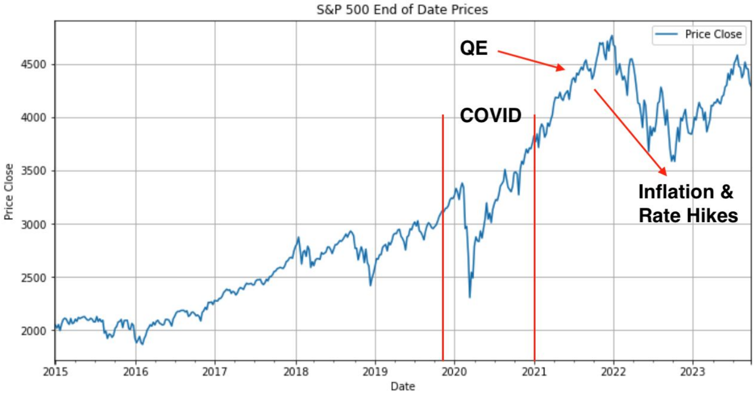
	Adj Close		Close		High		
	AAPL	SPY	AAPL	SPY	AAPL	SPY	AAPL
2021-10-04 13:30:00-04:00	138.751907	427.750092	138.751907	427.750092	139.220001	428.549988	138.535004
2021-10-04 14:30:00-04:00	138.585007	427.635010	138.585007	427.635010	139.320007	429.220001	138.509995
2021-10-04 15:30:00-04:00	139.169998	428.700012	139.169998	428.700012	139.229996	429.040009	138.470001
2021-10-05 09:30:00-04:00	139.869995	433.739990	139.869995	433.739990	140.399994	433.790009	139.410004

1. **Factor Formulation**
   b. Fundamental Factors
   c. Technical Factors
   d. Economic & Alternative Factors

1. **Data Cleaning (Stage 2)**
   b. Trimming Outliers
   c. Winsorizing Data
   d. Quality Checks & Validation







Bollinger bands

Main Idea: technical analysis tool that signal potential overbought or oversold conditions, aiding traders in identifying reversal points.

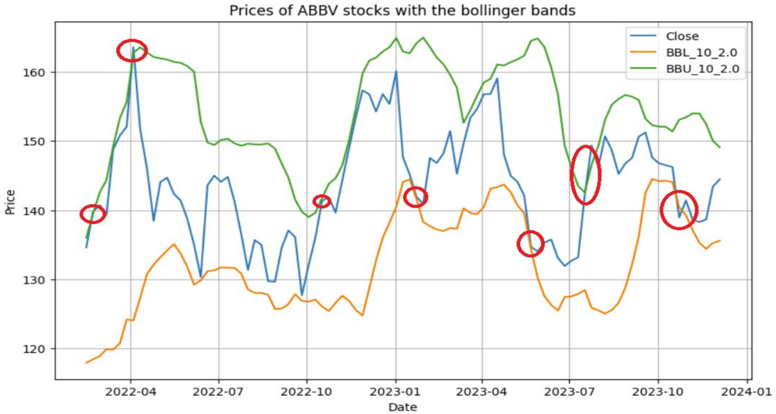
Implementation:

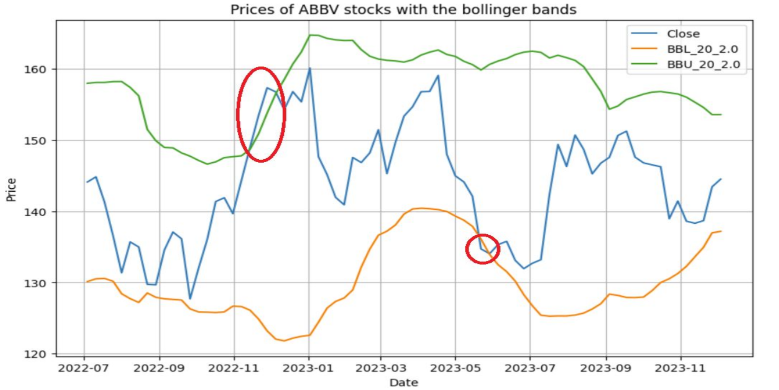
1. Get price history
2. Compute the x-day moving average where x is usually 20 or 50
3. Compute the lower band by subtracting 2 std to the moving average
4. Compute the upper band by adding 2 std to the moving average
5. Buy signal when the price is lower than the lower band and sell signal when the price is higher than the upper band

General framework:

1. **Bollinger bands:** Implement the bollinger bands for all the assets
2. **Create Bollinger variables:** that takes 1 if lower (higher) than the lower band (higher band), 0 otherwise.
3. **Backtesting:** On rolling weekly periods, regress next Friday's returns with this week's value. We obtained even higher significance using a continuous factor as opposed to (-1, +1, 0) setup:

$$F_{BB} = \frac{P(t) - P_{roll,average}}{\sigma_{roll,average}}$$





The SARIMAX Factor (pmdarima.arima.auto_arima)

Model Explanation: Stands for Seasonal AutoRegressive Integrated Moving Average with eXogenous factors model. Advanced version of ARIMA, incorporating seasonality and external variables.

ARIMA Usefulness:

1. Adapts to changing market conditions
2. Improves accuracy with external data
3. Handles non-stationarity data effectively
4. It can be used to fit other data than prices & returns

Seasonal Data:

1. Quarterly Earnings Reports
2. Holiday Sales Data (especially for retail/consumer companies & ETFs)

Exogenous:

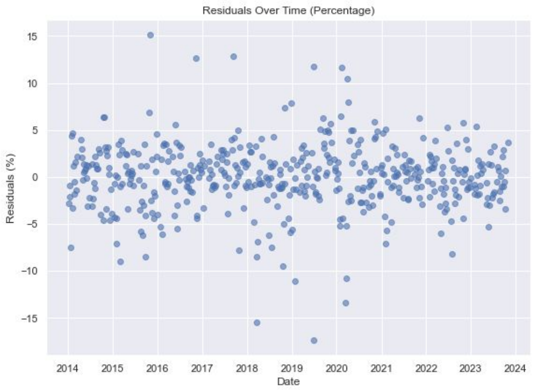
1. GDP Growth Rates
2. Interest Rates

Fitting AUTO_ARIMA:

1. Leverage Daily Data to more accurately describe the timeseries.
2. Convert to returns, given improved stability over prices
3. Automatically Selects relevant Auto-correlation Factors
4. Fit to predict the price over next week.

General Framework

1. **Verify Model Correctness:** Iterate through securities to test **auto_arima** in conjunction with more manual ARIMA models.
2. **Deduce Relevant Training Windows:**
   a. **Model Fitting:** Rolling Windows separated by 13 weeks
   b. **Model Training:** Windows of variable length (30 Weeks - 3 Years). Outputs: **Optimal Training Window**; **Modifications** based on available trading history (Crypto)
3. **Store Matrix of Predicted Returns, along with 21-day rolling volatility**
4. **Test Predicted vs Actual Prices**



The SARIMAX Factor (pmdarima.arima.auto_arima)

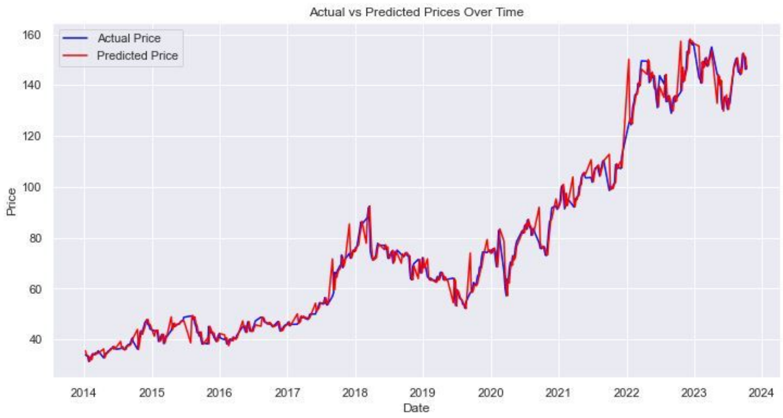
General Framework (Continued)

5. Smoothing Results & Eliminating Impact Days:

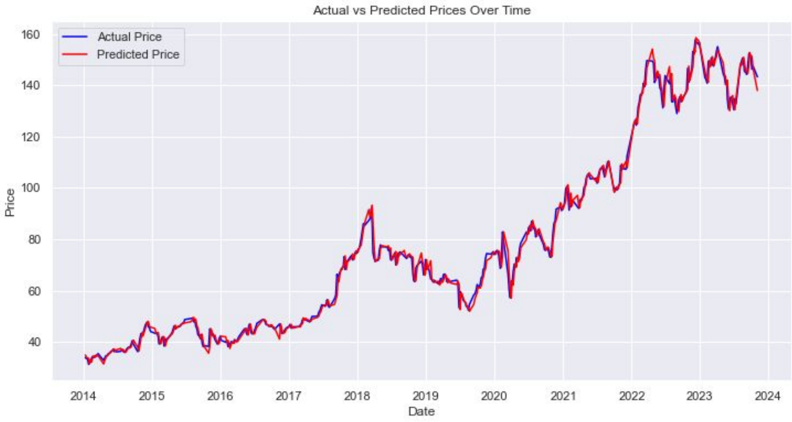
- **A. Interpolate data for non-trading days**
- **B. Eliminate impact days from dataset**
- **C. Include 12-month active T-Bill data**

Results display:

1. **Lower jumps in total**
2. **Smoother tracking of underlying security**
3. **Reduced Outliners overall**



Turning this



Into this

The SARIMAX Factor (pmdarima.arima.auto_arima)

Backtesting Framework:

1. ***Synthesize results from individual securities***
2. ***Create monitored factor:*** *Predicted price of next Friday over the rolling 21-day standard deviation of returns (iteratively):*

$$F_{ARIMA} = \frac{P_{predicted}}{\sigma_{21-day,rolling}}$$

1. ***Z-score factors annually:*** *Cross-sectionally compare the factor of each security over time to prepare data for the linear regression.*

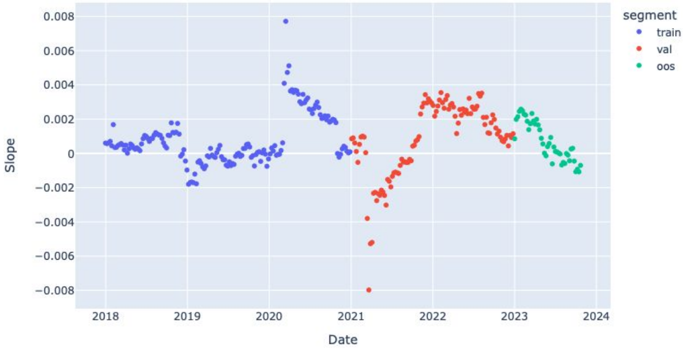
$$Z - score, F_{ARIMA} = \frac{F_{ARIMA,i} - \bar{F}_{ARIMA}}{SD}$$

1. ***Run linear regression, estimating statistical significance of results.***

1. ***Create a long-short portfolio, with "± w" on each security, with variable length on top & bottom winners, to perform sensitivity analysis on return over the benchmark index (SPX).***

Key Insights & Discussion:

1. ***Long-short weights prevent protect this portfolio against major drawdowns of the market***
2. ***Model sensitive to external factors that may show conflicting results across different time periods;***
   a. *Low interest rate environment of 2012 - 2019 adversely affects model performance for post-COVID era.*
   b. *Similar sensitivity to unemployment; Low unemployment signals strong market performance pre-COVID, but pushes markets downwards signalling strong economy & stubborn inflation (eliminated).*



Embedding Analyst Recommendations

- **Data**: Analysts recommendations from Refinitiv (idea from *Ratings Changes, Ratings Levels, and the Predictive Value of Analysts' Recommendations*)
- Attribute a score to each recommendation and combine them all with an average score. The lower the score the more likely the stock's price will go up.
- **Backtesting** : Regress next week's return with current week's score

SUMMARY GRID

RECOMMENDATION SUMMARY

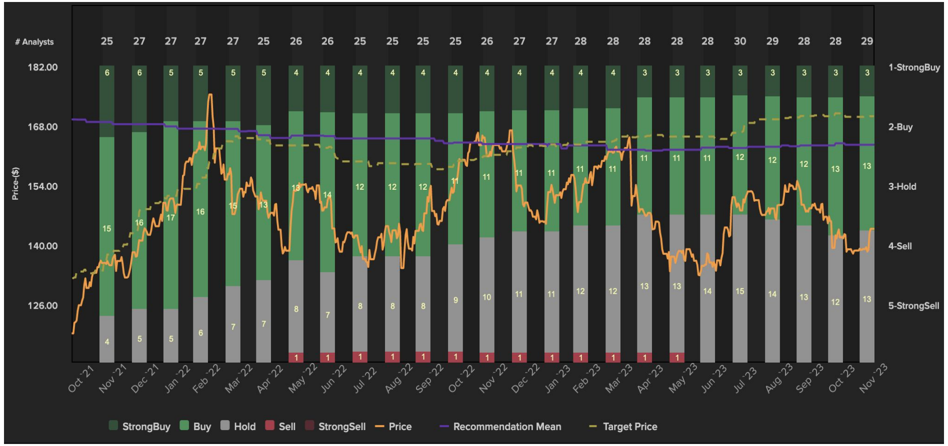
	Analysts Per level			
	04-Sep-2023	04-Oct-2023	04-Nov-2023	Current
1-StrongBuy	3	3	3	3
2-Buy	12	12	13	13
3-Hold	14	13	12	13
4-Sell	-	-	-	-
5-StrongSell	-	-	-	-
Rec Mean	**2.4**	**2.4**	**2.3**	**2.3**

Source : Refinitiv (ticker : ABBV)

Embedding Analyst Recommendations

- **Pros**:
  - Easy to use
  - Leverages the work of analysts

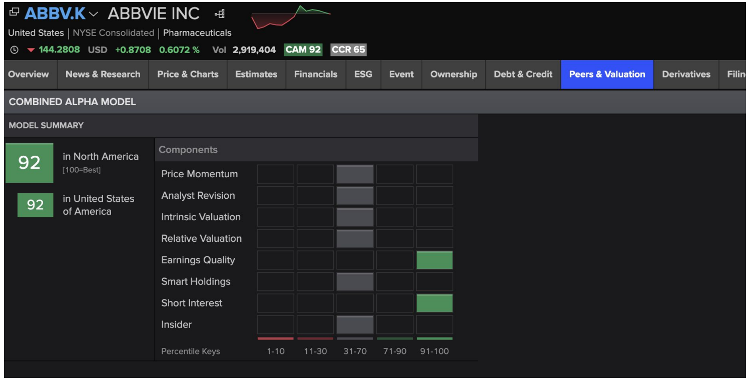
- **Cons**:
  - Very little statistical significance
  - Not suited for weekly decision making
  - Falls into the herding mentality bias
  - Depends on the stock's coverage

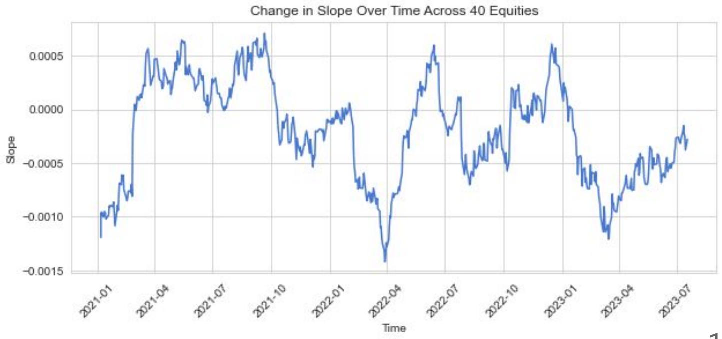


Source : Refinitiv (ticker : ABBV)

Starmine Reuters Alpha Model

- **Main Idea:** Leverage Refinitiv's Starmine Alpha model
  - Ranks stock based on various metrics
    - Price Momentum
    - Analyst Valuation
    - Intrinsic Valuation
    - Earnings Quality
    - Short Interest
  - Normalizes Value to be between 1-100
- **Pros**:
  - Leverage industry insights
  - Completely internal to Reuters
  - Individual Components Statistically Significant
- **Cons**:
  - Correlated with other signals
    - Causing Multicollinearity
  - Overall not statistically significant





Combination of Models (Black Litterman Model)

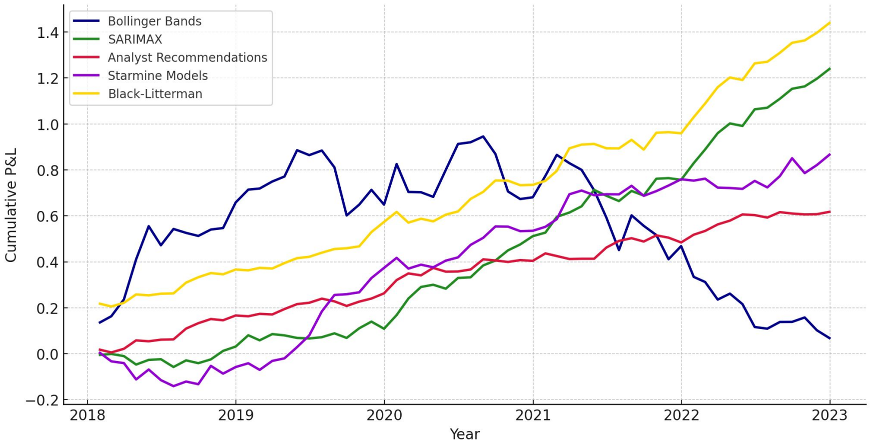
$$\Pi = \lambda \Sigma w_{mkt}$$

- Π the Implied Excess Equilibrium Return Vector
- λ the risk aversion coefficient
- Σ the covariance matrix of excess returns
- w_{mkt} the market capitalization weight of the assets

$$E[R] = \left[(\tau\Sigma)^{-1} + P'\Omega^{-1}P\right]^{-1}\left[(\tau\Sigma)^{-1}\Pi + P'\Omega^{-1}Q\right]$$

- E[R] the new (posterior) Combined Return Vector
- Σ the covariance matrix of excess returns
- P matrix that identifies the assets involved in the views
- τ scalar
- E[R] the new (posterior) Combined Return Vector
- Ω diagonal covariance matrix of error terms from the expressed views (uncertainty in each view)
- Q the View Vector

Combination of Models (Black Litterman Model)

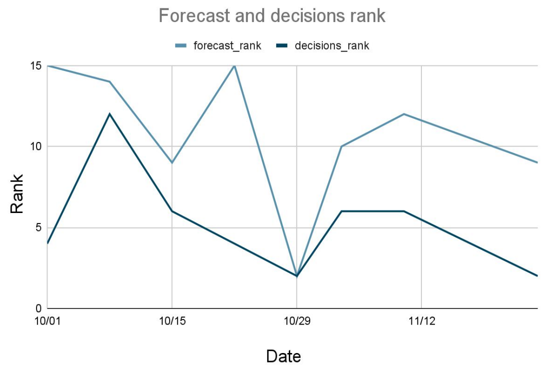


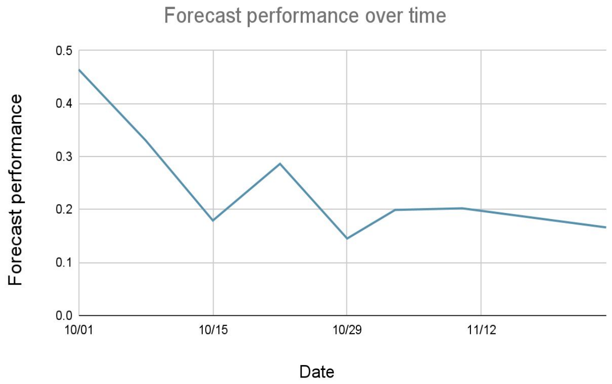
- **Pros**:
  - Incorporates **Investor Views** and confidence levels.
  - **Flexibility**: Can be adapted to a wide range of asset classes and investment scenarios.
  - Combines **Multiple Strategies**: Effectively integrates various investment strategies.

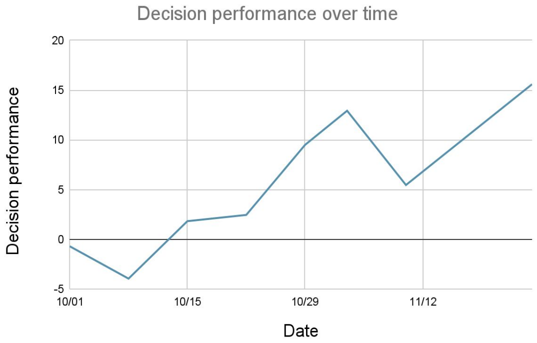
- **Cons**:
  - More **complex to implement** and understand than simpler models like Bollinger Bands or SARIMAX.
  - **Dependent on Quality of Inputs** (investor's views and the covariance matrix).
  - There's a **risk of overfitting** the model to past data.

The Fund’s Performance

- ***As we incorporated more robust features over time, we performed better, improving forecasting & decisions performance.***
- ***Last week’s poor performance largely affected by the introduction of new factors, presenting an outlier & adversely affecting the semester’s results.***







Thank you & Disclaimer

The Citadel Spinoffs

Disclaimer: The findings of this presentation should under no circumstances serve to solicit financial decisions, and are solely produced for the class "*IEOR 4576 - Data Driven Methods in Finance*" at Columbia University.

Thank you for your attention!

Questions?