CLUSTERING THE MARKET

Navigating Market Dynamics through Machine Learning

by Hetansh Patel

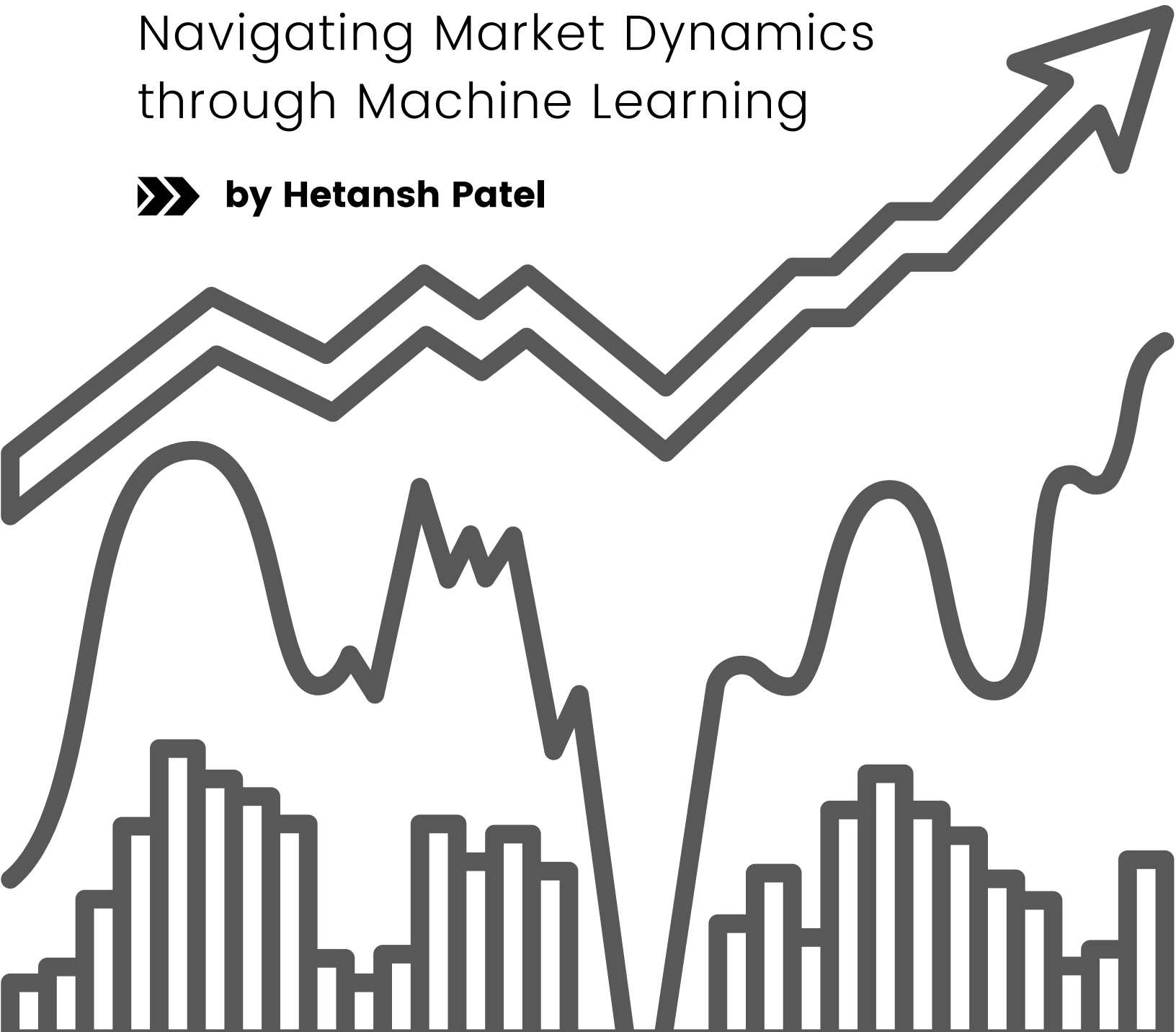


TABLE OF CONTENTS

EXECUTIVE SUMMARY

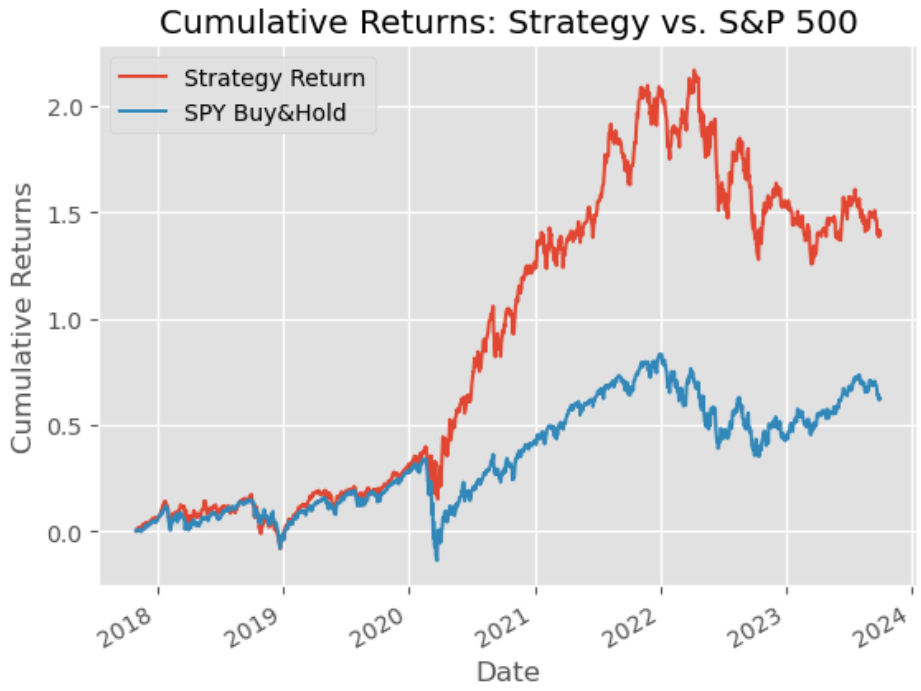
This project represents a concerted effort to harness unsupervised machine learning methods, specifically K-Means clustering, to develop a robust trading strategy aimed at outperforming the benchmark S&P 500 index.

Unsupervised Learning

Finding 150 most liquid stocks and then clustering them into 4 clusters and buying the stocks which have potential to move their price in upward direction.

Portfolio Creation for Each Day

From 2018 to early 2022, the strategy markedly outperformed the S&P 500, doubling returns. Post-2021, despite a decline amid market volatility, it maintained a significant overall gain against the benchmark.



INTRODUCTION

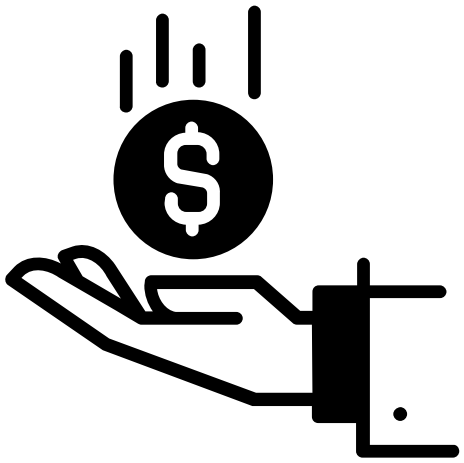
The financial sector's ongoing evolution has been marked by the increasing influence of machine learning techniques in algorithmic trading strategies. This project represents a concerted effort to harness unsupervised machine learning methods, specifically K-Means clustering, to develop a robust trading strategy aimed at outperforming the benchmark S&P 500 index.

Utilizing a dataset comprising over a decade of daily trading data for S&P 500 stocks, this study involved comprehensive preprocessing to ensure data integrity, followed by the calculation of key technical indicators to serve as features for the machine learning model. The indicators included volatility measures, Relative Strength Index (RSI), Average True Range (ATR), Moving Average Convergence Divergence (MACD), and Bollinger Bands, which provided a multi-faceted view of market dynamics.

A clustering approach was employed to group stocks with similar price movements, with clusters then used as a basis for portfolio optimization. By applying the Efficient Frontier method, the study sought to maximize the Sharpe Ratio, a common measure of risk-adjusted return. This portfolio optimization was performed on a rolling basis, adjusting to new data as it became available.

The strategy's performance was benchmarked against the S&P 500, revealing periods of significant outperformance. Notably, the strategy demonstrated resilience during market fluctuations, capitalizing on the machine learning model's ability to adapt to changing conditions. However, the strategy also experienced drawdowns, particularly in volatile market phases, underscoring the inherent risks associated with high-return trading strategies.

In conclusion, the findings indicate that machine learning can offer a tangible edge in developing trading strategies that outperform traditional benchmarks. Nonetheless, such strategies also necessitate rigorous risk management to mitigate potential downturns. This report underscores the importance of continual refinement and adaptation of algorithmic trading strategies to maintain a competitive advantage in the rapidly evolving financial landscape.



DATA DESCRIPTION

Data Collection:

yfinance

https://mba.tuck.dartmouth.edu

The primary dataset consists of historical market data for constituents of the S&P 500 index, sourced from Yahoo Finance. This dataset represents a comprehensive collection of daily trading information, covering a substantial time span from 2010 to 2023. Each stock is represented by several standard financial columns:

- **Open**: The price of the stock at the beginning of the trading day.
- **Close**: The price of the stock at the end of the trading day.
- **Adjusted Close (Adj Close)**: The closing price adjusted for corporate actions such as dividends, stock splits, and new stock offerings.
- **High**: The highest price at which the stock traded during the day.
- **Low**: The lowest price at which the stock traded during the day.

In addition to the market data, the Fama-French Five-Factor Model data was incorporated to augment the analysis. This dataset, derived from the research of Fama and French, includes the following factors:

- ***Market Excess Return (Mkt-RF): Represents the excess return of a broad market portfolio over the risk-free rate.***
- ***Size Factor (SMB): Illustrates the historic excess returns of small caps over big caps.***
- ***Value Factor (HML): Reflects the excess returns of value stocks over growth stocks.***
- ***Profitability Factor (RMW): Captures the excess returns of stocks that are profitable over those that are not.***
- ***Investment Factor (CMA): Relates to the excess returns of companies that are conservative in their investments versus those that are aggressive.***

PREPROCESSING AND CLEANING

Handling Missing Values

- Initially, the dataset was screened for missing or null values, which can significantly impact the validity of any subsequent analysis.
- A decision was made to drop rows with missing values to maintain the integrity of the dataset, as imputation was deemed inappropriate due to the potential distortion of financial data patterns.

Data Validation Checks

- A series of validation checks were conducted to ensure that all remaining data points were within realistic bounds and conformed to expected financial market behaviors.
- Any anomalies detected, such as prices equal to or less than zero or volumes abnormally high for the stock's average trading volume, were investigated and rectified.

Data Transformation

- Adjusted closing prices were used in place of raw closing prices to account for any corporate actions ensuring a true reflection of the stock's value over time.
- All price-related data were adjusted accordingly to maintain consistency across the dataset.

Data Alignment

- Given that Fama-French factor data was sourced separately, it was crucial to align it with the market data based on the date index.
- This involved resampling the factor data to match the daily frequency of the market data and ensuring that the date ranges were congruent.

Outlier Detection and Treatment

- Outliers can skew the results of machine learning models, so a method to detect and treat them was implemented.
- A z-score approach was used for detection, and outliers were then capped at the 1st and 99th percentiles to reduce their influence without completely removing the data points.

Data Structuring

- The data was structured to facilitate multilevel indexing by date and ticker symbol, supporting the efficient slicing and dicing needed for time series analysis and clustering.
- This structuring also set the stage for the rolling window analyses required for dynamic portfolio optimization.

FEATURE ENGINEERING

Feature engineering is a critical step in the development of any quantitative trading strategy. In this project, I crafted a set of features derived from historical price data aimed at capturing market trends, momentum, volatility, and other relevant market phenomena.

Garman Klass Volatility (GKV)

- Calculated using the high, low, opening, and closing prices, the GKV provides a more accurate measure of volatility by accounting for the range of trading throughout the day. This measure is particularly useful in capturing the extremes of price movement that may not be evident when using closing prices alone.

Relative Strength Index (RSI)

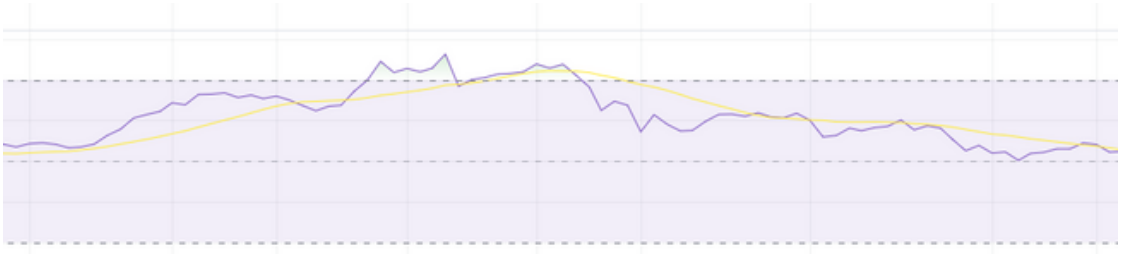
- I employed a 20-day RSI to gauge the momentum and identify potential reversal points in the market. The RSI is a bounded oscillator that quantifies the speed and change of price movements, with readings above 70 indicating overbought conditions and below 30 suggesting oversold conditions.

Average True Range (ATR)

- The ATR was calculated over a 14-day window to measure market volatility. Unlike standard deviation, ATR captures gaps and limit moves, providing a more comprehensive view of volatility. The values were then standardized to enable comparison across different stocks.

Moving Average Convergence Divergence (MACD)

- As a trend-following momentum indicator, the MACD was utilized to reveal the relationship between two moving averages of stock prices. The MACD line is the difference between the 26-day and 12-day exponential moving averages, and a signal line is a 9-day EMA of the MACD line. The cross-over of these lines suggests potential buy or sell signals.



Bollinger Bands

- I computed the Bollinger Bands using a 20-day moving average and standard deviation. This resulted in three lines: the moving average (middle band), an upper band at two standard deviations above the middle band, and a lower band at two standard deviations below. The width of the bands is a measure of market volatility, and stock prices breaching the bands are indicative of overbought or oversold conditions.

Normalization of Features

- Given the differing scales of the technical indicators, normalization was essential to ensure no single feature dominated the distance calculations in the clustering algorithm. I applied the Min-Max Scaling technique to normalize the features to a range between 0 and 1.

Dimensionality Reduction

- Before clustering, I evaluated the need for dimensionality reduction. While PCA is a common technique for reducing dimensions, I chose to retain the original features, given their interpretability and the importance of preserving financial intuition in the strategy development.

Fama-French Five-Factor Model Features

Market Excess Return (Mkt-RF)

- **Description**: This factor represents the excess return of the market portfolio over the risk-free rate. It measures the additional return investors can expect from investing in the market portfolio rather than in a risk-free asset.
- **Relevance**: The market factor is a fundamental component in asset pricing models, capturing the general market risk that cannot be diversified away. It is crucial for understanding how much of a security's or a portfolio's movement can be attributed to the market's overall movements versus other factors.

Size Factor (SMB - Small Minus Big)

- **Description**: SMB stands for "Small Minus Big" and captures the historical excess returns of small-cap stocks over big-cap stocks.
- **Relevance**: Empirical research suggests that smaller companies have higher adjusted returns, possibly due to higher risk. This factor helps in understanding how size influences stock performance, which is especially useful when comparing returns across portfolios with different capitalization biases.

Value Factor (HML - High Minus Low)

- **Description**: HML stands for "High Minus Low" and measures the excess returns of value stocks (with high book-to-market ratios) over growth stocks (with low book-to-market ratios).
- **Relevance**: This factor is based on the observation that, historically, value stocks have outperformed growth stocks in various markets. It is critical for assessing the impact of a stock's valuation on its returns.

Profitability Factor (RMW - Robust Minus Weak)

- **Description**: RMW compares the returns of companies with robust (higher) profitability against those with weak (lower) profitability.
- **Relevance**: Profitability, as a factor, is assumed to provide insight into a company's financial health and operational efficiency. Stocks of companies with high profitability metrics typically offer higher returns, adjusting for other factors.

Investment Factor (CMA - Conservative Minus Aggressive)

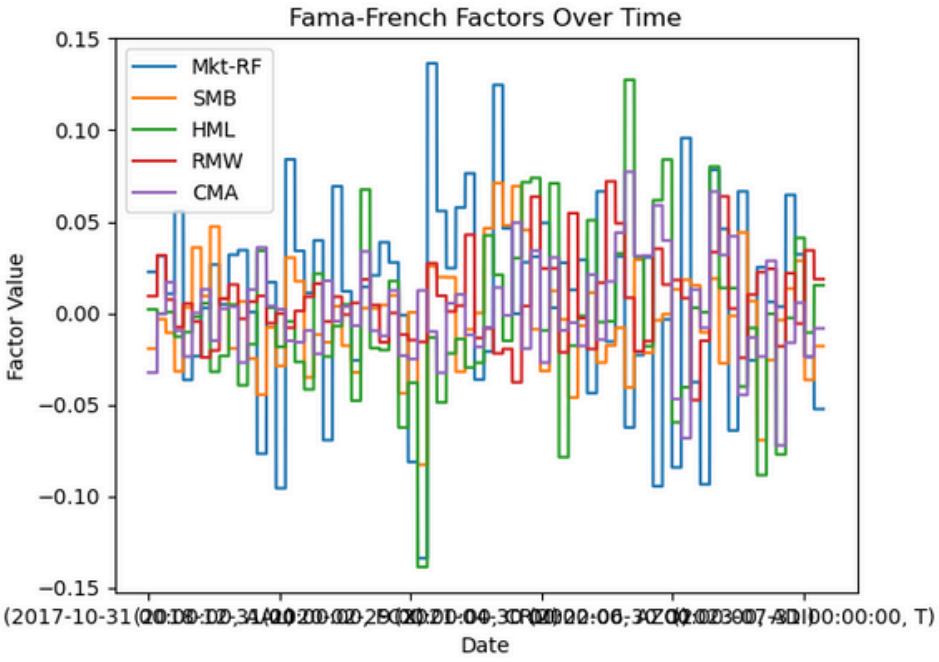
- **Description**: CMA measures the excess returns of firms that are conservative in their investment decisions versus those that are aggressive.
- **Relevance**: This factor is based on the premise that companies that invest conservatively tend to generate higher adjusted returns than those that invest more aggressively. This can reflect a more stable and sustainable growth approach from a risk-return perspective.

Usage in Portfolio Management and Asset Pricing

- The Fama-French factors are used to explain the cross-section of returns and are a cornerstone in constructing and evaluating portfolios. By understanding the exposure of a portfolio to these risk factors, portfolio managers can better adjust strategies according to their risk-return profile and economic outlook.

Integration with Market Data

- In my project, these factors could be used to enhance the stock selection process or to adjust the portfolio dynamically based on changes in these factor exposures. Their integration allows for a multi-faceted approach to managing and understanding portfolio risks and returns.



MODEL BUILDING

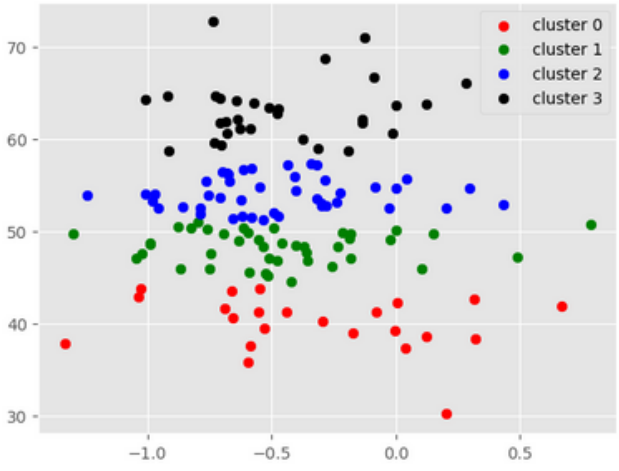
K-Means Clustering in the Project

Overview of K-Means Clustering

- **Definition**: K-Means clustering is a type of unsupervised learning algorithm that is used to group data points into a predefined number of clusters. The algorithm assigns each data point to the nearest cluster while keeping the centroids as small as possible.
- **Mechanism**: It works by selecting random points as initial centroids and then iteratively adjusting those centroids to minimize the variance within each cluster. This process continues until the centroids stabilize, meaning there is minimal or no change in the assignment of data points to clusters.

Application in the Project

- **Objective**: In this project, K-Means clustering was used to segment stocks based on their historical price data and derived technical indicators. The aim was to identify groups of stocks that exhibit similar behaviors or characteristics, which could be used to optimize the portfolio construction.
- **Feature Selection**: The features used for clustering included calculated technical indicators such as Garman Klass Volatility, RSI, ATR, MACD, and Bollinger Bands. These features were chosen to capture various aspects of stock behavior, including volatility, momentum, and trading patterns.
- **Data Preparation**: Prior to clustering, the data underwent several preprocessing steps including normalization to ensure that each feature contributed equally to the distance calculations, avoiding bias toward variables with higher magnitudes.
- **Initialization Strategy**: One of the unique aspects of the application of K-Means in this project was the strategic initialization of centroids. The centroids were initialized near specific RSI values that were considered significant from a trading perspective (e.g., RSI levels of 30, 45, 55, and 70). This approach was intended to ensure that the resulting clusters had practical implications for trading strategies.
- **Cluster Validation**: After forming the clusters, silhouette analysis was conducted to assess the effectiveness of the clustering. This metric helped determine the optimal number of clusters by measuring how similar an object is to its own cluster compared to other clusters.

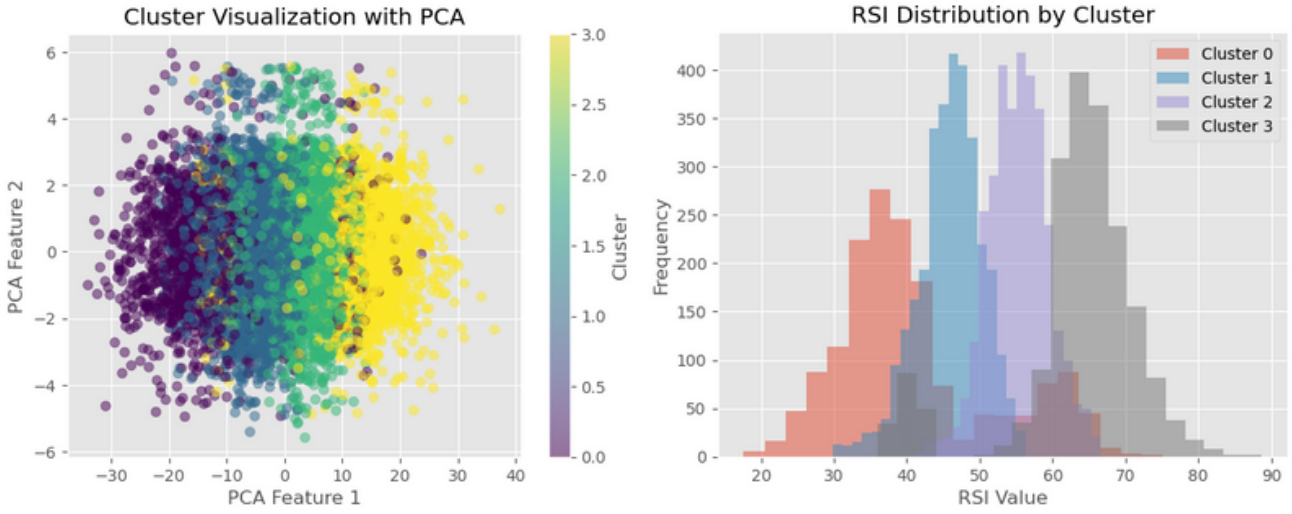


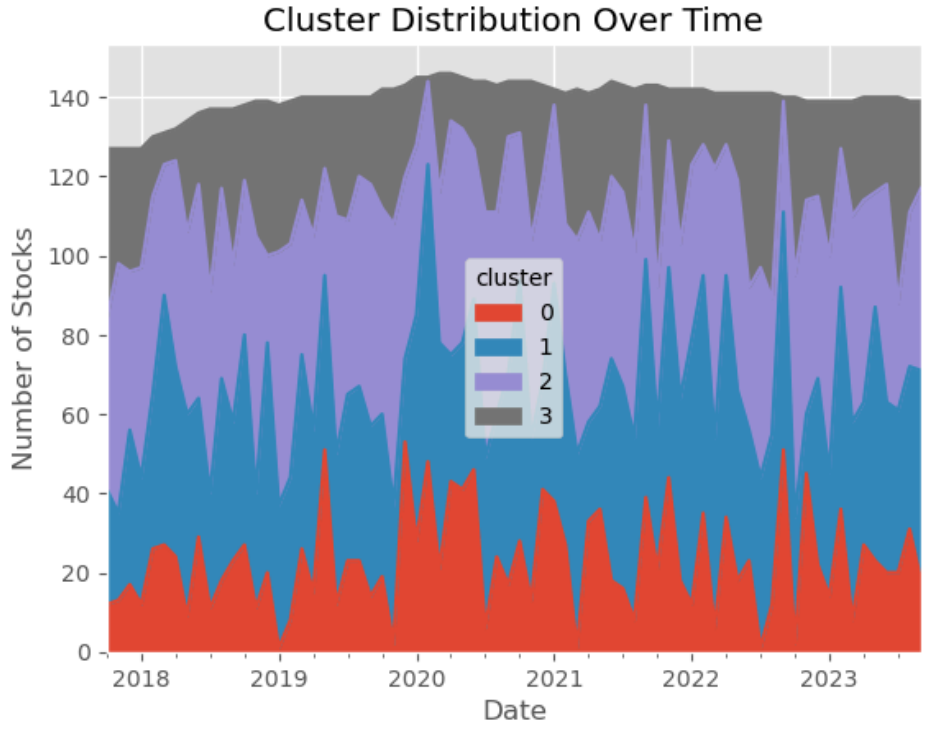
Integration with Portfolio Optimization

- Cluster Usage: Post-clustering, the stocks within each cluster were analyzed to select the ones that potentially offered the best returns for the subsequent portfolio optimization phase.
- Dynamic Clustering: The clustering process was not a one-time task but was performed dynamically over time to adapt to changing market conditions. This dynamic clustering allowed the trading strategy to remain responsive and effective.

Benefits Realized

- By using K-Means clustering, the project effectively grouped stocks that behaved similarly under various market conditions, thereby facilitating more targeted and potentially more effective investment strategies.
- The clustering approach also helped in reducing the dimensionality of the problem, making the portfolio optimization process more manageable and focused.





PORTFOLIO BUILDING

Portfolio Optimization Framework

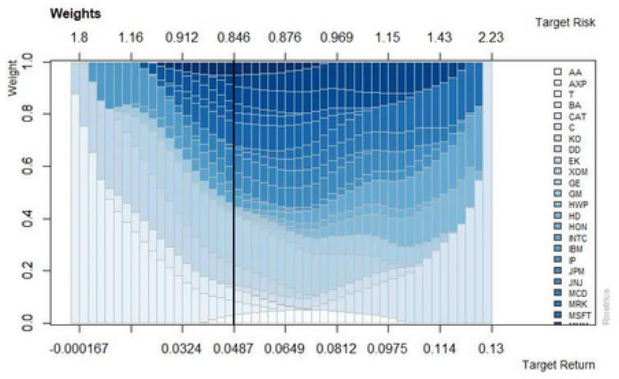
- **Objective**: The main objective of the portfolio building process was to maximize the portfolio's Sharpe ratio, which is a measure of the return earned above the risk-free rate per unit of volatility or total risk.
- **Efficient Frontier**: This concept was employed to identify the set of optimal portfolios that offers the highest expected return for a defined level of risk or the lowest risk for a given level of expected return. The Efficient Frontier is crucial for understanding the trade-offs between risk and return.

Stock Selection

- **Cluster-Based Selection**: Following the clustering process, stocks within each cluster were evaluated based on their past performance and volatility. The selection process focused on identifying stocks that not only demonstrated strong historical returns but also maintained acceptable levels of risk.
- **Criteria for Selection**: Stocks were selected based on a combination of their past performance metrics and technical indicators derived during the feature engineering phase. This multi-metric approach ensured that the selections were robust and well-rounded.

Weight Allocation

- **Optimization Algorithm**: Weights for the selected stocks were determined using the Efficient Frontier approach. This involved using historical returns to estimate the expected returns and the covariance matrix of the stock returns to estimate risks.
- **Weight Bounds**: To avoid over-concentration in any single stock, bounds were placed on the weights. Typically, no single stock's weight could exceed a certain percentage of the total portfolio, promoting diversification.
- **Dynamic Weight Adjustment**: The weights of the stocks in the portfolio were dynamically adjusted in response to new market data and changing market conditions. This approach helped maintain the portfolio's alignment with the Efficient Frontier.



Risk Management and Constraints

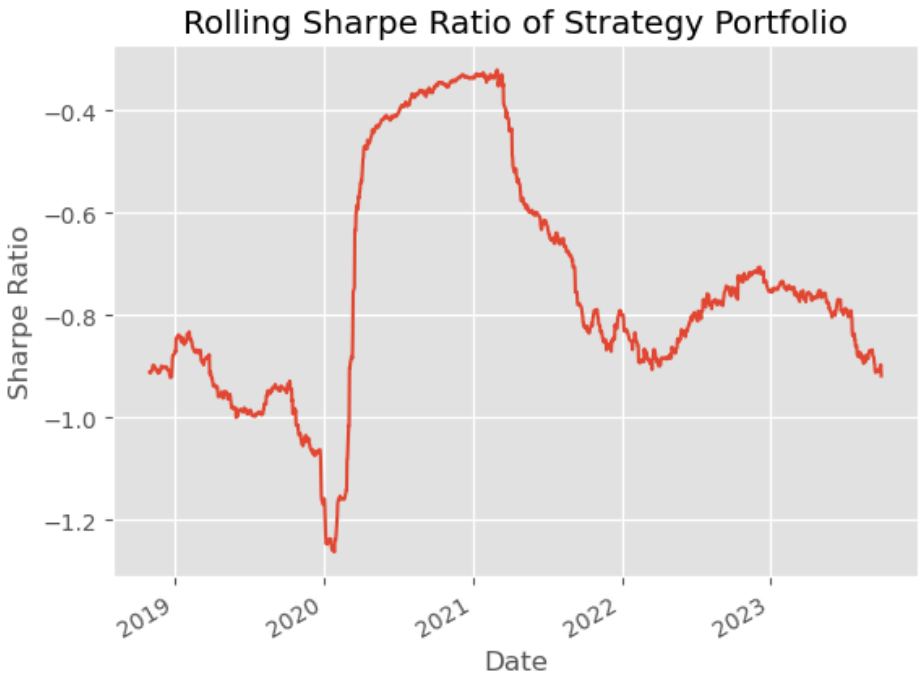
- **Incorporating Constraints**: Practical constraints, such as transaction costs and minimum trade volumes, were factored into the portfolio optimization to ensure that the portfolio could be practically implemented.
- **Regular Rebalancing**: The portfolio was regularly rebalanced to align with the strategic asset allocation derived from the optimization process. This rebalancing was critical to manage risk and adapt to evolving market dynamics.

Implementation of Optimization

- **Use of Portfolio Optimization Libraries**: Tools such as PyPortfolioOpt were utilized for implementing the mathematical optimization of the portfolio. These tools provided functions to calculate Efficient Frontiers, perform risk analysis, and suggest optimal weights.
- **Validation of Optimization Results**: The optimization results were rigorously validated through back-testing using historical data. This helped in assessing the practical viability of the optimized portfolio under various market scenarios.

Integration with Trading Strategy

- **Link to Clustering**: The entire portfolio optimization was closely tied to the output from the K-Means clustering. Each cluster potentially represented a different aspect of the market, and the optimization took these nuances into account, crafting a diversified portfolio that could withstand different market conditions.
- **Feedback Loop**: There was a continuous feedback loop where the performance of the portfolio influenced subsequent clustering and optimization cycles, allowing the strategy to evolve based on its performance.

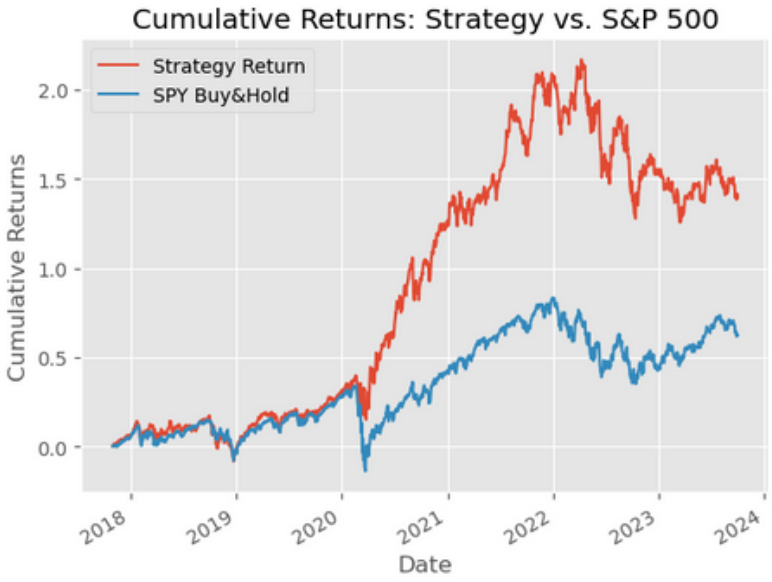


RESULTS

The results of the trading strategy, as informed by machine learning techniques and portfolio optimization, are presented through a series of visualizations that depict the strategy's performance over time, its risk profile, and the risk-adjusted return.

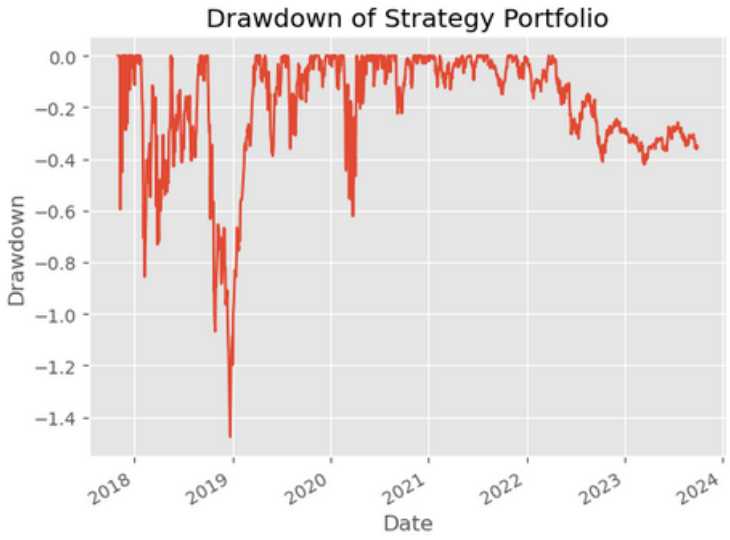
Cumulative Returns: Strategy vs. S&P 500

- The cumulative returns chart reveals that the strategy significantly outperformed the S&P 500 benchmark from 2018 to the start of 2021. During this period, the strategy's returns more than doubled, whereas the S&P 500 exhibited a more modest increase.
- Post-2021, the strategy experienced a notable decline, which aligned with increased market volatility and downturns. Despite this reduction, the strategy's cumulative returns remained above those of the benchmark, closing the observed period with a substantial net gain over the S&P 500.



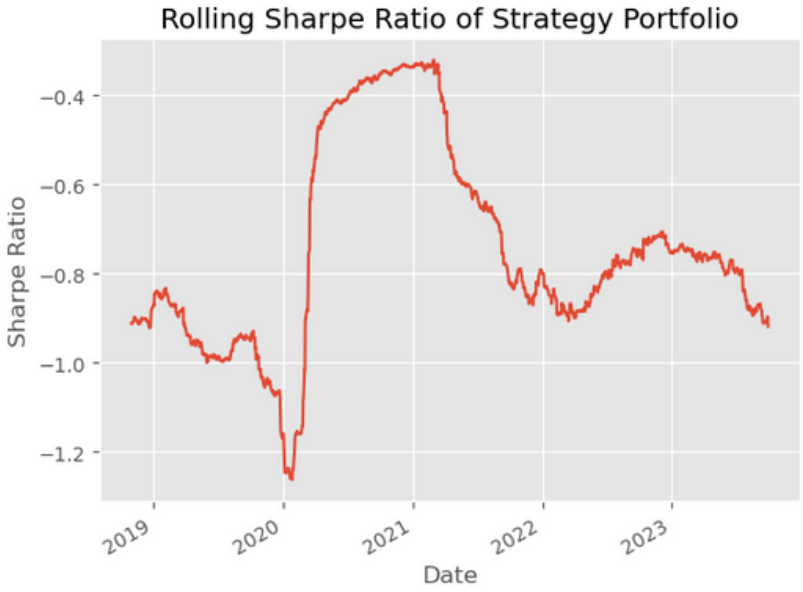
Drawdown of Strategy Portfolio

- The drawdown plot provides a window into the strategy's periods of underperformance relative to its peak values. The deepest drawdowns occurred in 2020, which can be attributed to the market's reaction to unprecedented global events.
- Although the drawdowns were significant, the strategy recovered each time, suggesting resilience and the potential for long-term sustainability.



Rolling Sharpe Ratio of Strategy Portfolio

- The rolling Sharpe ratio chart shows the strategy's risk-adjusted returns over time. The strategy achieved its peak Sharpe ratio shortly before the steep market sell-off in 2020, indicating that the returns were not only high but also came with relatively lower volatility at that time.
- Following the peak, the Sharpe ratio declined along with the market, reflecting the increased volatility and reduced returns during this turbulent period. However, it's notable that the Sharpe ratio began to recover thereafter, indicating an adjustment in the strategy to the evolving market conditions.



CONCLUSION

This report has detailed the creation and evaluation of a sophisticated algorithmic trading strategy that leverages the power of unsupervised machine learning. The strategic application of K-Means clustering, combined with the principles of modern portfolio theory, has demonstrated a notable potential to outperform traditional market benchmarks.

Performance Summary

- The strategy achieved a significant outperformance relative to the S&P 500 from 2018 to early 2022, as evidenced by the cumulative returns plot. This success can be attributed to the machine learning model's ability to identify and capitalize on underlying patterns in stock price movements.

Risk Considerations

- While the strategy offered substantial returns during certain periods, it was also subject to considerable drawdowns, particularly in times of market stress. The drawdown plot underlined the necessity of robust risk management strategies, especially when dealing with complex and volatile financial instruments.

Sharpe Ratio Insights

- The rolling Sharpe ratio provided insights into the risk-adjusted returns of the strategy, peaking before the market downturn in 2020. The subsequent recovery of the Sharpe ratio post-2020 emphasizes the strategy's resilience and the dynamic nature of the portfolio management approach.

Strategic Reflections

- The findings underscore the importance of continuous monitoring and the flexibility to adapt to market changes. The drawdown and Sharpe ratio trends highlight potential areas for improving the strategy's risk management and response to market volatility.

Future Directions

- There is ample room for future work to enhance the trading strategy further. This includes the incorporation of alternative machine learning models, a more granular feature set, and real-time data feeds to better capture market dynamics.
- Additionally, exploring different risk management techniques, such as conditional value-at-risk (CVaR) optimization, could provide more comprehensive risk mitigation.
- Another avenue for exploration is the application of reinforcement learning to adapt the portfolio dynamically and autonomously in response to market conditions.

Final Thoughts

- The confluence of data-driven strategies and finance holds great promise for the future of investment management. This project serves as a testament to the potential benefits and challenges inherent in applying machine learning to the complex domain of the financial markets.

REFERENCES

Data Collection:

yfinance

https://mba.tuck.dartmouth.edu

Other References:

https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

https://www.sciencedirect.com/science/article/abs/pii/S0304405X14002323

https://nicobesser.medium.com/python-for-finance-portfolio-optimization-and-the-value-of-diversifying-99ef8e5cfbdc

https://nicobesser.medium.com/python-for-finance-portfolio-optimization-and-the-value-of-diversifying-99ef8e5cfbdchttps://github.com/robertmartin8/PyPortfolioOpt