S&P500 Stock Analysis

Swami Venkatesan

10/19/2023

Outline

- S&P 500 Overview
- Data Processing
- Sector-wise Analysis
- Stock Performance Distribution and Labels
- Stock Clustering by Trend
- Combined Model for Improved Classification
- Summary and Areas of Improvement

Overview

- Standard and Poor's 500 (S&P500) is a stock market index that tracks 500 leading publicly traded companies in the U.S
- The S&P 500 uses a market-cap weighting method, giving a higher percentage allocation to companies with the largest market capitalizations
- These stocks belong to one of the 11 different sectors

Data Mining and Wrangling

- The latest list of S&P500 companies is downloaded from a scraper
- 2 year daily market data (OCHLV) for all stocks is then retrieved using Yahoo Finance API
- To account for any sudden price changes, the following steps were taken
  - Average of daily High and Low price to reduce intraday volatility
  - Simple moving average of 15 days (SMA10) to reduce day to day volatility
- SMA10 prices were used for 2-year stock performance calculations and labeling

Sector Composition and Weights

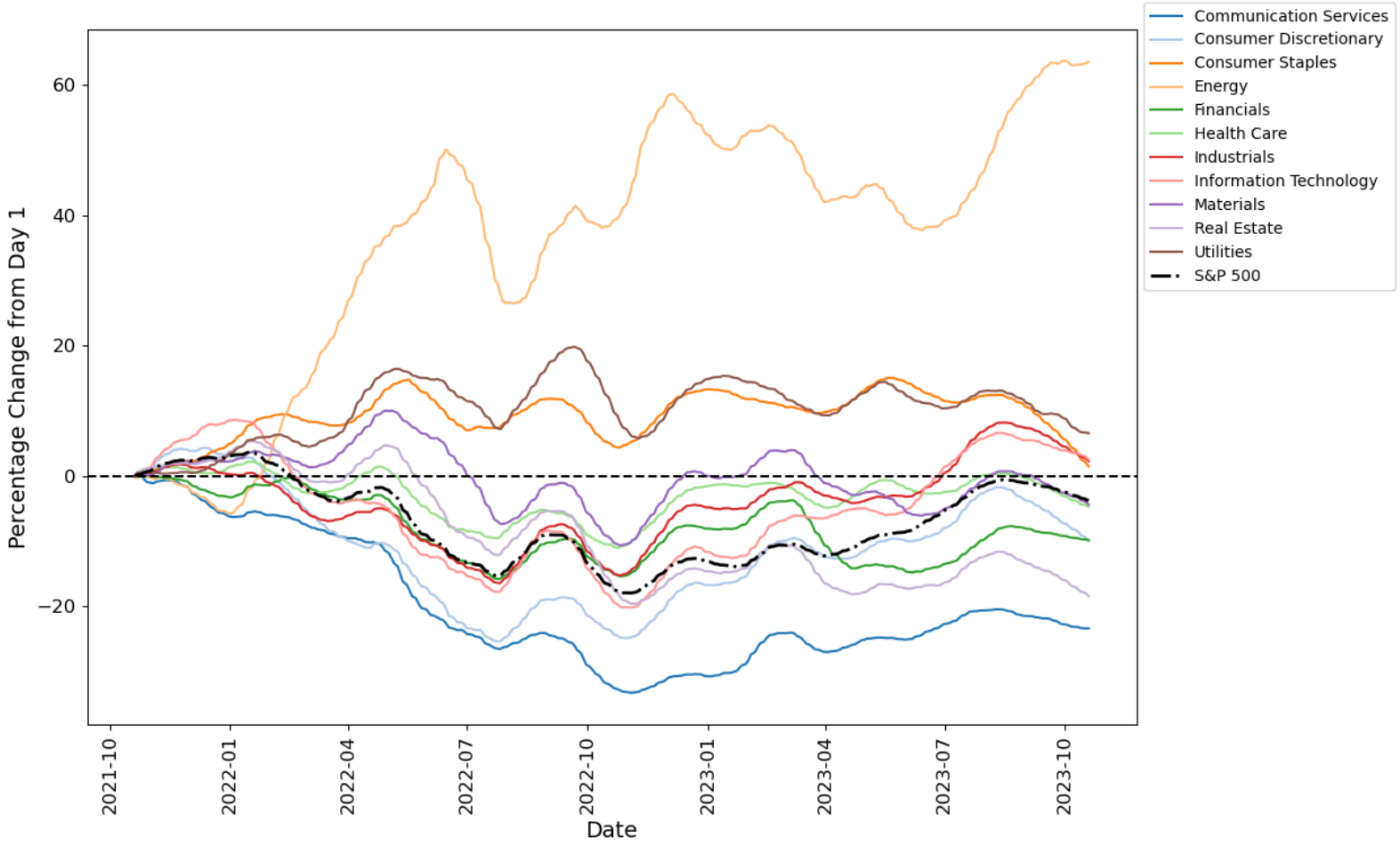
Sector	Sum of Stock Weights	Number of Stocks in Sector	Percentage of Stocks in SP500
Communication Services	9%	22	4%
Consumer Discretionary	11%	53	11%
Consumer Staples	6%	38	8%
Energy	5%	23	5%
Financials	11%	72	14%
Health Care	13%	64	13%
Industrials	8%	77	15%
Information Technology	28%	64	13%
Materials	2%	29	6%
Real Estate	2%	31	6%
Utilities	2%	30	6%

- Stocks in IT sector have the maximum weightage on S&P500
- IT, Industrial, Financial and Healthcare have the

2 Year Performance by Sector

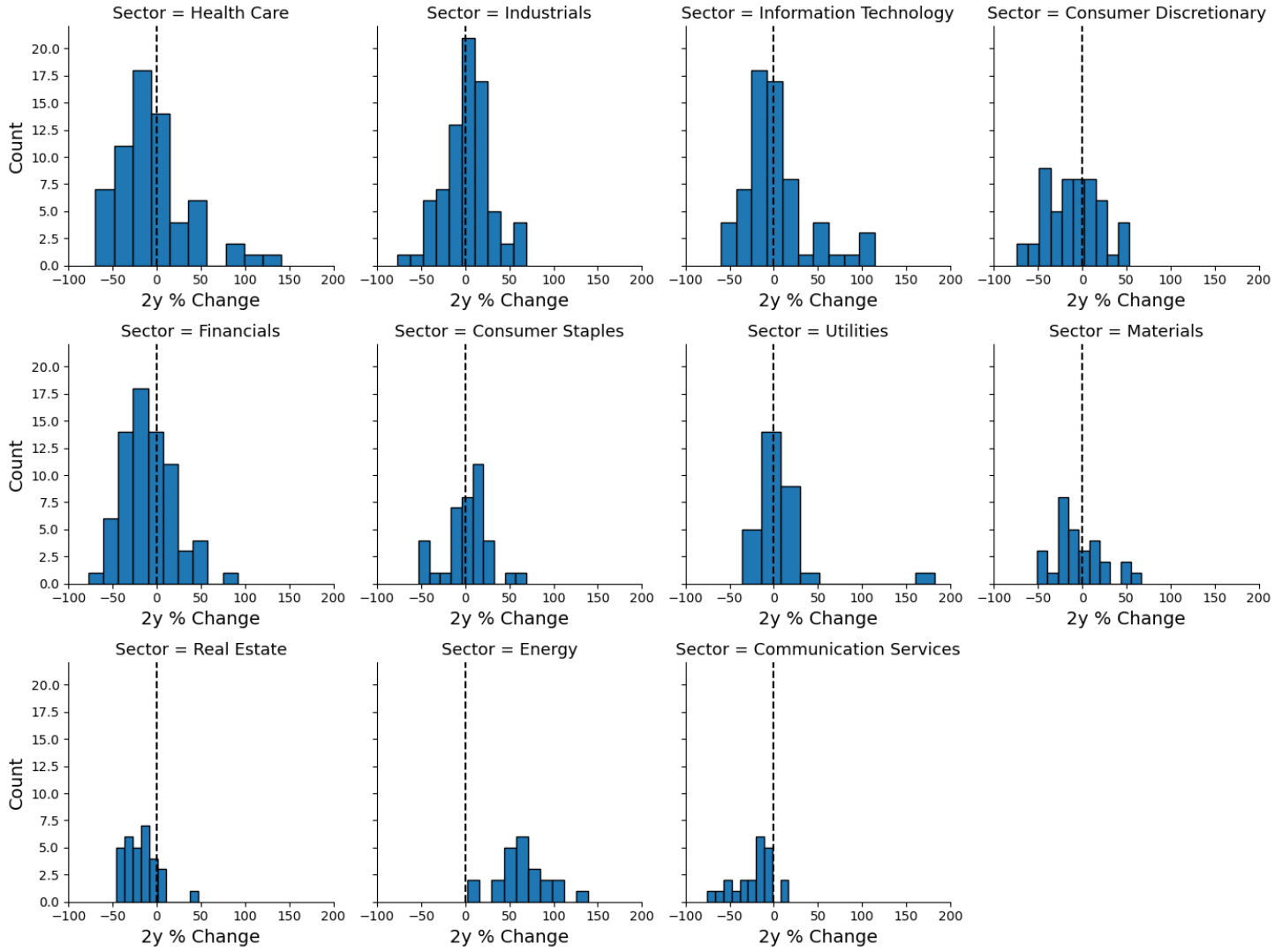
- Energy sector stocks show the highest gains
- Comm Service stocks show the lowest gain
- S&P index seems to follow higher weightage sectors – IT, healthcare and Financials

Sector	2y % Change	
	Mean	Median
Communication Services	-23%	-17%
Consumer Discretionary	-10%	-10%
Consumer Staples	2%	7%
Energy	64%	62%
Financials	-9%	-12%
Health Care	-3%	-12%
Industrials	3%	2%
Information Technology	3%	-5%
Materials	-4%	-11%
Real Estate	-18%	-18%
Utilities	8%	-1%



Sector Wise Stock 2y Performance Distribution

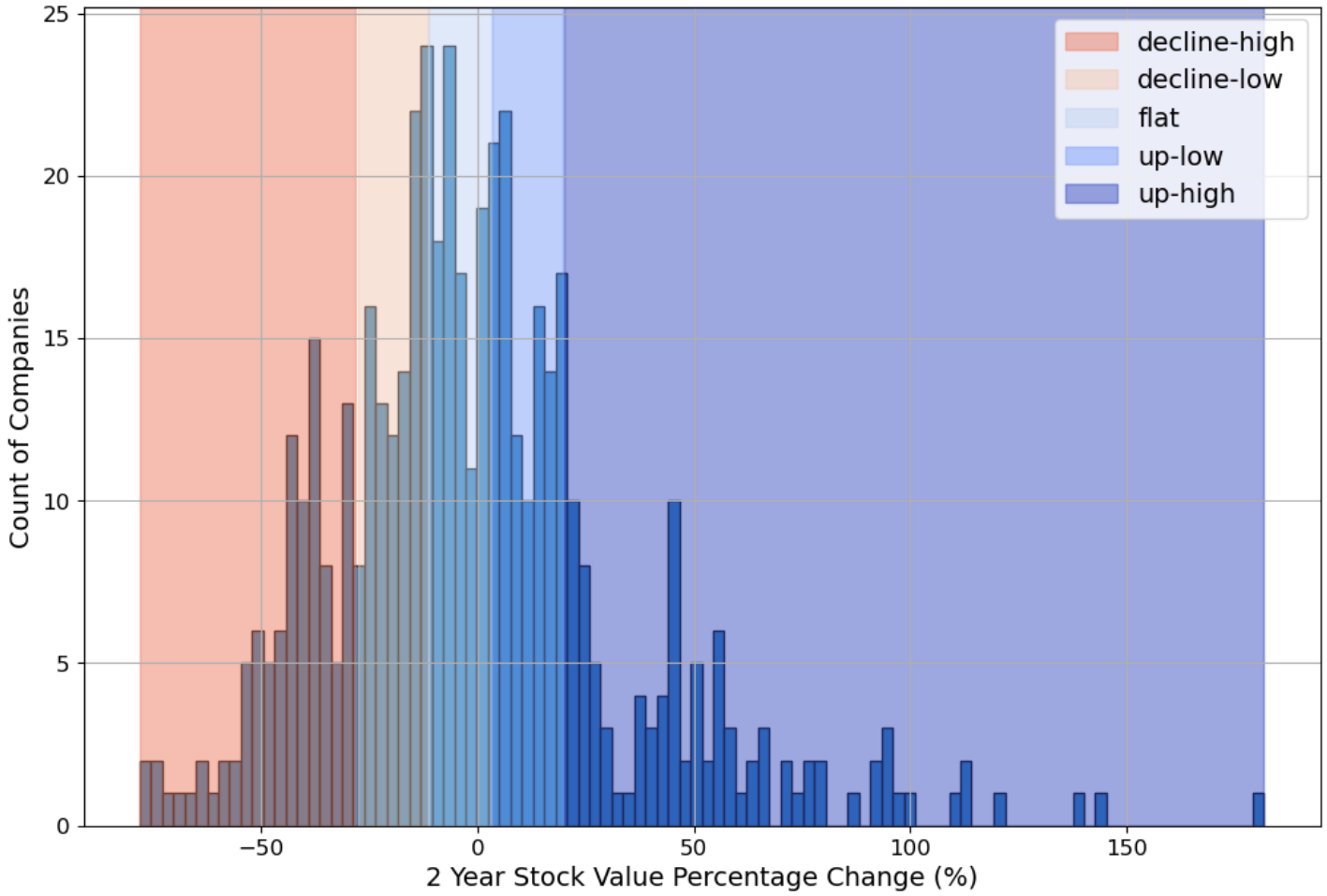
- Plot shows distribution of stocks vs 2-year percentage change
- All stocks in Energy sector show positive performance (>0%)
- Majority of the stocks in Real Estate and Comm. Services show negative performance
- All other sectors are distributed between positive and negative



Stock Performance Labels

- The percentage change between the price of the stock from initial price(Oct 18, 2021) to end of 2-year price was calculated
- The stocks were then automatically assigned to 5 different labels based on their 2-year performance (% change)
  - The binning was done by splitting the stocks into 5 quantiles

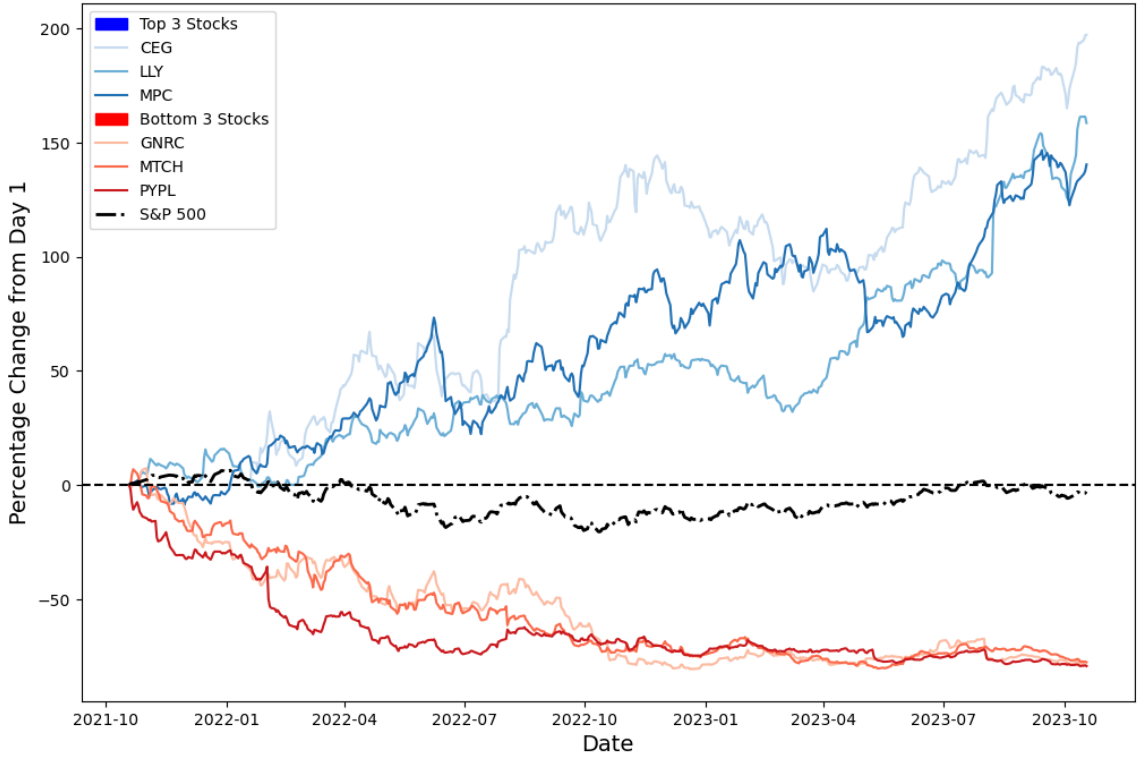
Performance Label	2y % Change Bins	
	Min	Max
decline-high	-80%	-31%
decline-low	-31%	-14%
flat	-14%	1%
up-low	2%	18%
up-high	18%	165%



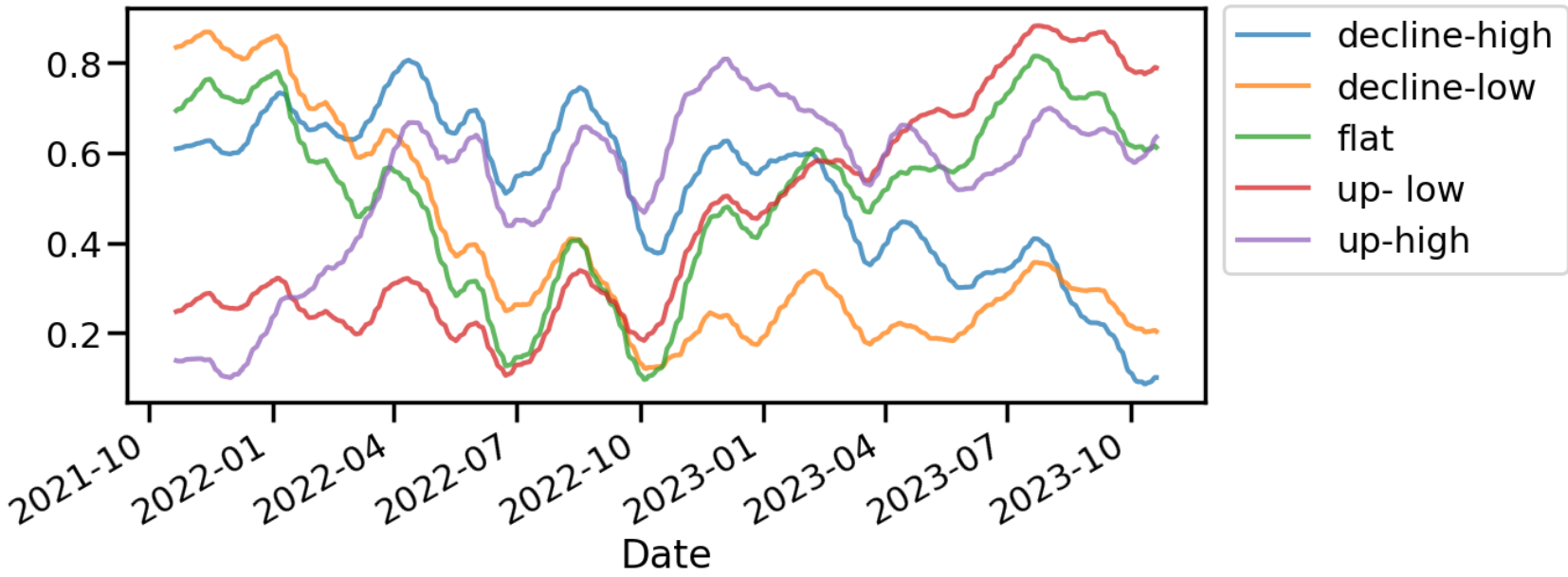
Top 3 and Bottom 3 Performing Stocks

- Top 3 and bottom 3 performing stock is plotted
- The daily change is calculated as percentage change from Day 1 (Oct 18,2021) prices
- The 6 stocks belong to different sectors

Ticker	$\text{Price}_{\text{Oct 18 2021}}$	$\text{Price}_{\text{Oct 18 2023}}$	2y % Change	Sector	Weightage
CEG	$39.25	$110.57	182	Utilities	0.10
LLY	$234.79	$572.23	144	Health Care	1.34
MPC	$63.51	$151.17	138	Energy	0.16
MTCH	$159.85	$40.12	-75	Comm. Services	0.03
GNRC	$468.32	$107.43	-77	Industrials	0.02
PYPL	$271.36	$59.37	-78	Financials	0.17



Time Series Clustering

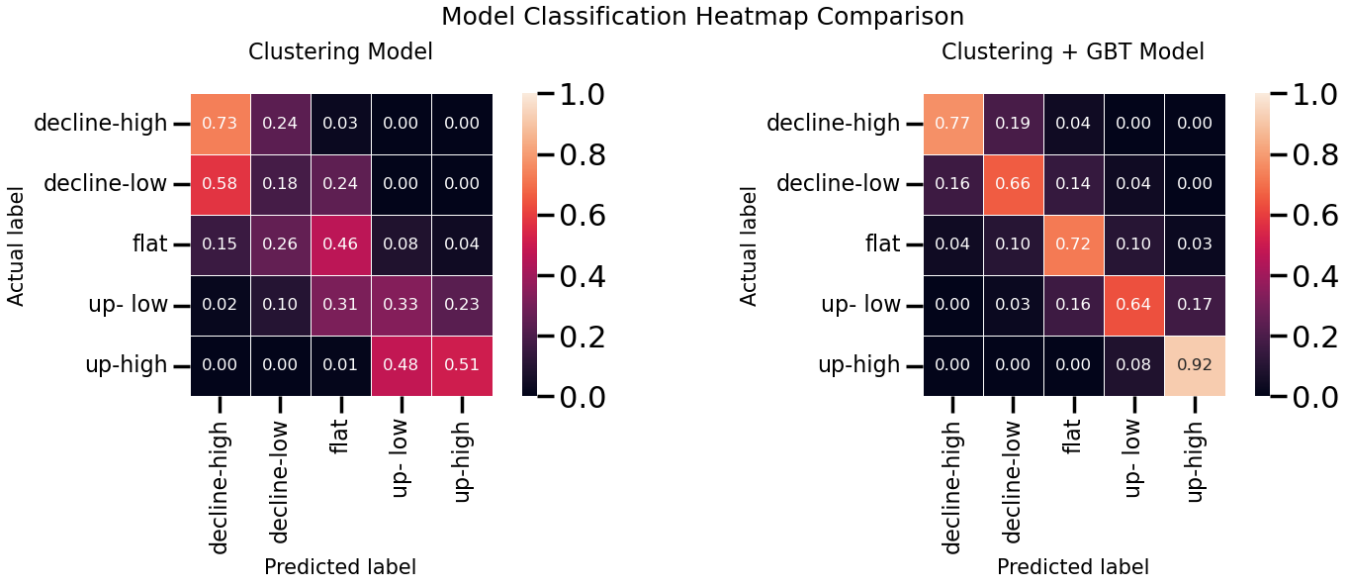


- To segment the stocks based on similarities in their trend, time series clustering method was used
  - Euclidean distance was used as the series is smoothened to remove short term volatility
- Time series clustering was done on the 2 year market data (freq = daily)
  - Each stock's timeseries was normalized between 0 and 1
- 5 clusters were used to match with binned labels earlier identified
- The clusters were then labeled similar to the 2y performance label based on the 2y % change

Model Metrics

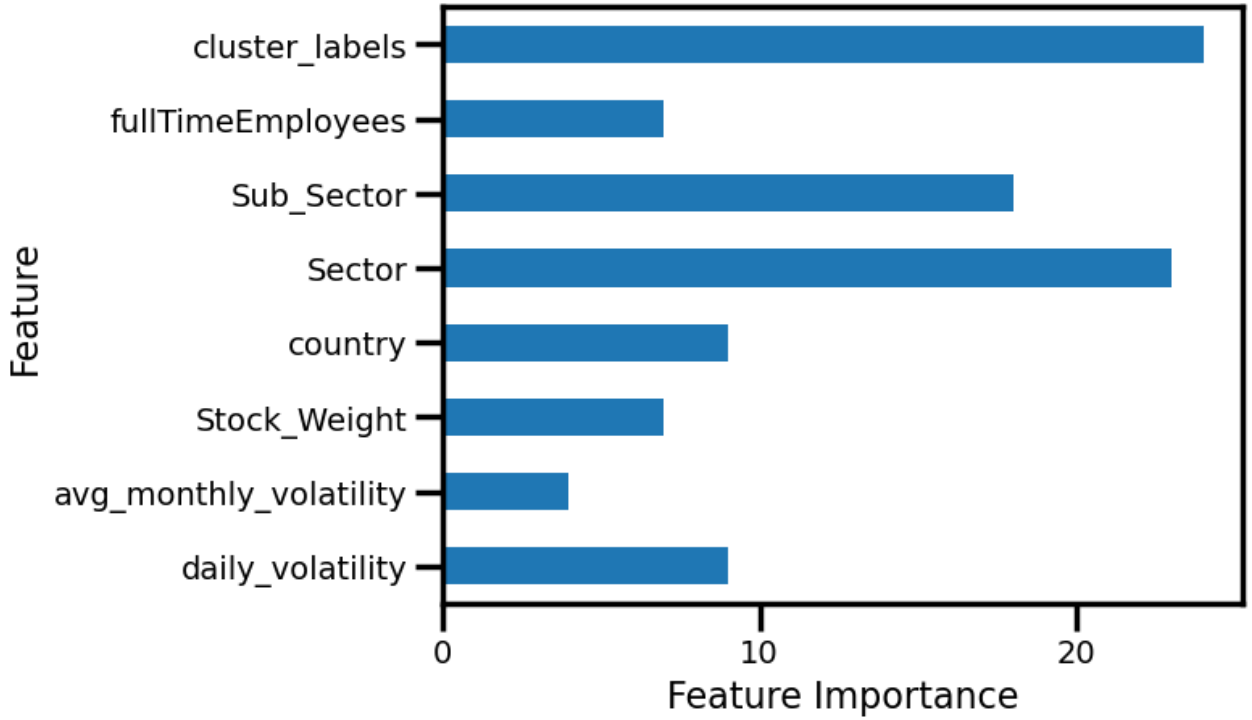
- All the stocks were then assigned to the cluster using the time series clustering algorithm
  - Accuracy of the clustering algorithm is ~ 45%
- Gradient Boosted Tree (GBT) model was also developed using the clustering labels as one of the feature (Clustering + GBT)
- The model was trained on 80% of data and showed 62% accuracy on the remaining unseen 20% data
  - Overall accuracy of the combined Clustering + GBT = 74%
  - Significant improvement in classifying mid labels – decline-low, flat and up-low

Model	Accuracy
Clustering	45%
Clustering + GBT	74%



Feature Importance

- The importance index from Catboost (GBT) model of each feature is plotted
- Cluster labels from time series clustering show the highest importance, followed by Sector and Sub-Sector



Summary

- An analysis on S&P500 stocks and sector performance was done
- Stocks were labelled based on their 2y performance
- Time series clustering was used to cluster the stocks based on their market trends and the quality of the clustering was evaluated against the 2y performance
- The accuracy of the performance labelling was improved using the cluster labels as a feature along with other stock related features – sector, sub-sector,price volatility etc.
- The model can be further improved by adding more static and dynamic features
  - Historical relative sector strength
  - Market Cap
  - Earnings Trend
  - Balance Sheet information