Clustering S&P 500 Stocks

Group 34: Eric Shi, Tahmid Washy, Esha Parikh, & Yiming Yuan

Background

- Stocks are currently categorized by industry
- These industries are broad, such as ‘Finance’ and ‘Industrials’
- Industries are not always indicative of stock performance

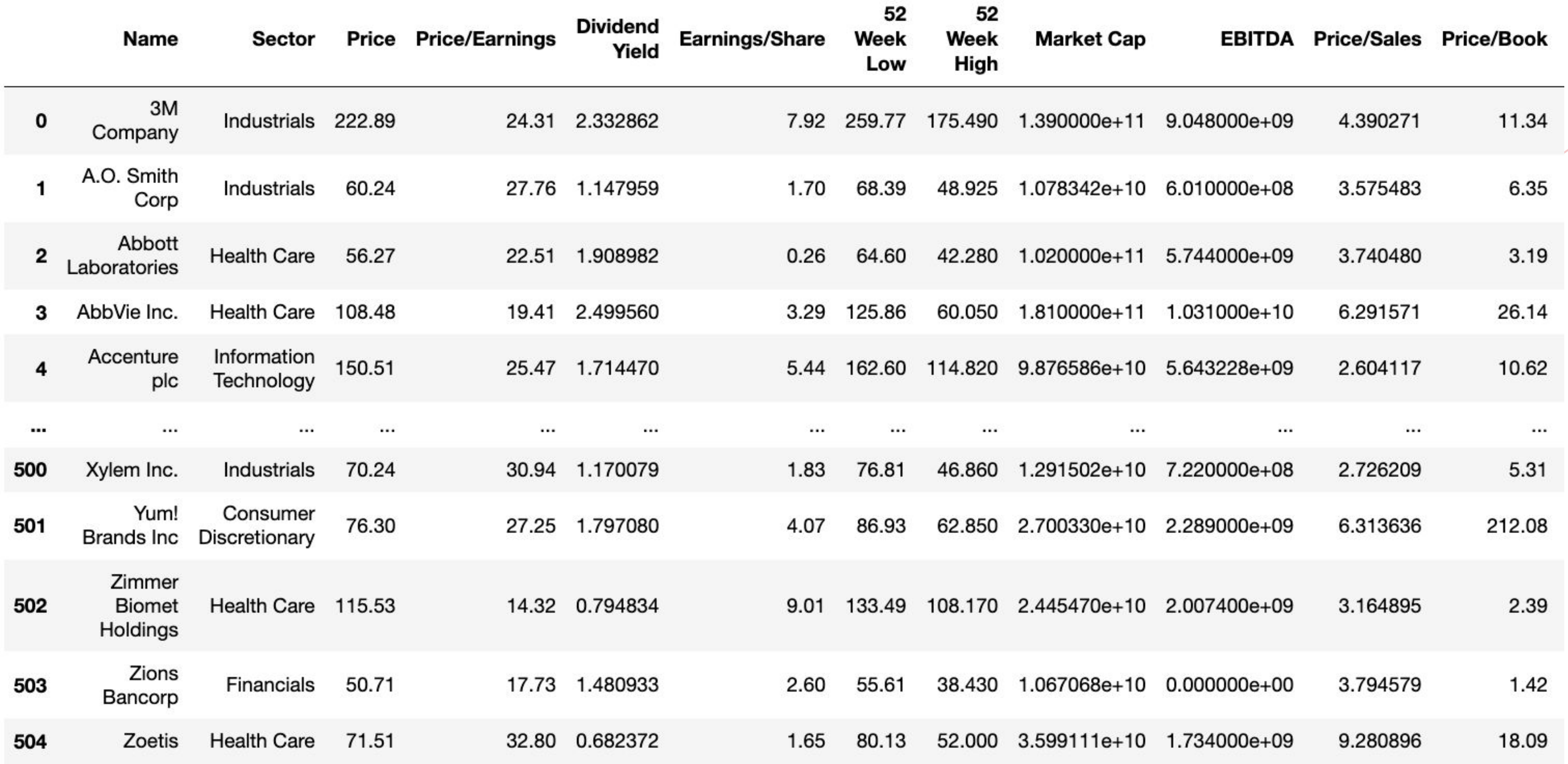
The Big Question

Is there a way to classify equities that is more indicative of their stock movement?

e.g. ‘Long-term’ vs. ‘Short-term’ stocks

Our Dataset

‘constituents - financials.csv’



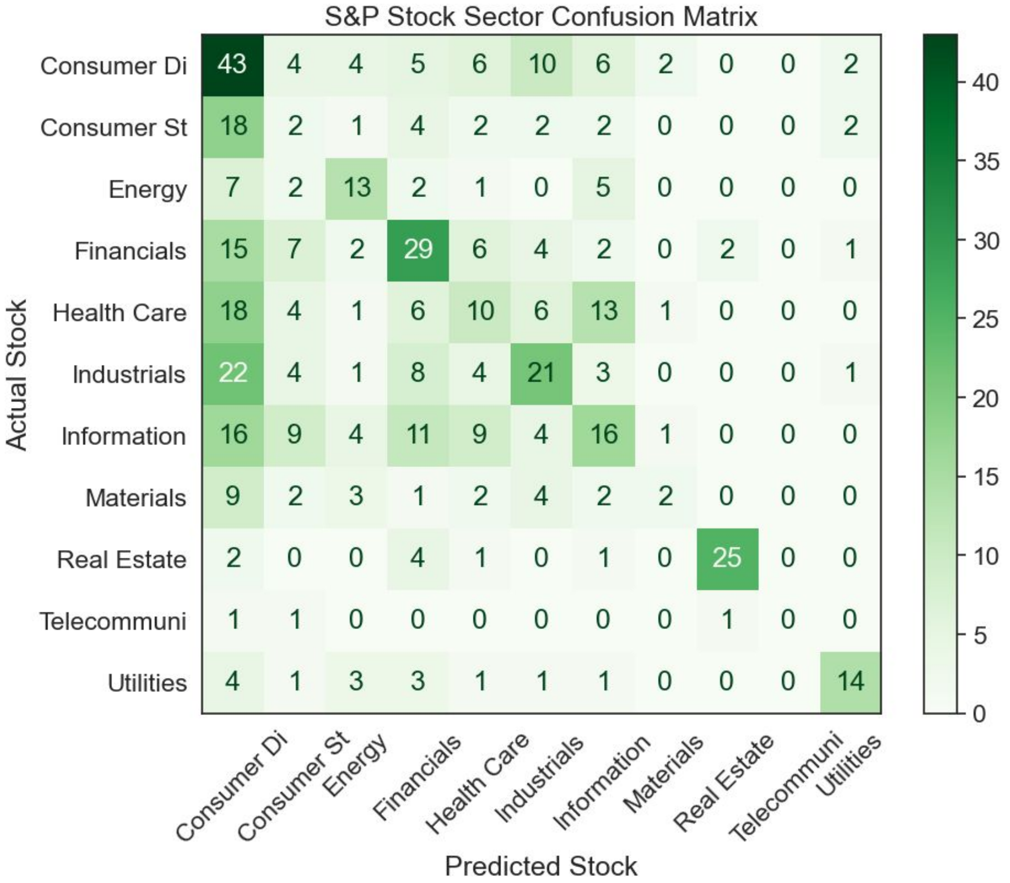
	Name	Sector	Price	Price/Earnings	Dividend Yield	Earnings/Share	52 Week Low	52 Week High	Market Cap	EBITDA	Price/Sales	Price/Book
0	3M Company	Industrials	222.89	24.31	2.332862	7.92	259.77	175.490	1.390000e+11	9.048000e+09	4.390271	11.34
1	A.O. Smith Corp	Industrials	60.24	27.76	1.147959	1.70	68.39	48.925	1.078342e+10	6.010000e+08	3.575483	6.35
2	Abbott Laboratories	Health Care	56.27	22.51	1.908982	0.26	64.60	42.280	1.020000e+11	5.744000e+09	3.740480	3.19
3	AbbVie Inc.	Health Care	108.48	19.41	2.499560	3.29	125.86	60.050	1.810000e+11	1.031000e+10	6.291571	26.14
4	Accenture plc	Information Technology	150.51	25.47	1.714470	5.44	162.60	114.820	9.876586e+10	5.643228e+09	2.604117	10.62
...	...	...	...	...	...	...	...	...	...	...	...	...
500	Xylem Inc.	Industrials	70.24	30.94	1.170079	1.83	76.81	46.860	1.291502e+10	7.220000e+08	2.726209	5.31
501	Yum! Brands Inc	Consumer Discretionary	76.30	27.25	1.797080	4.07	86.93	62.850	2.700330e+10	2.289000e+09	6.313636	212.08
502	Zimmer Biomet Holdings	Health Care	115.53	14.32	0.794834	9.01	133.49	108.170	2.445470e+10	2.007400e+09	3.164895	2.39
503	Zions Bancorp	Financials	50.71	17.73	1.480933	2.60	55.61	38.430	1.067068e+10	0.000000e+00	3.794579	1.42
504	Zoetis	Health Care	71.51	32.80	0.682372	1.65	80.13	52.000	3.599111e+10	1.734000e+09	9.280896	18.09

Methodology

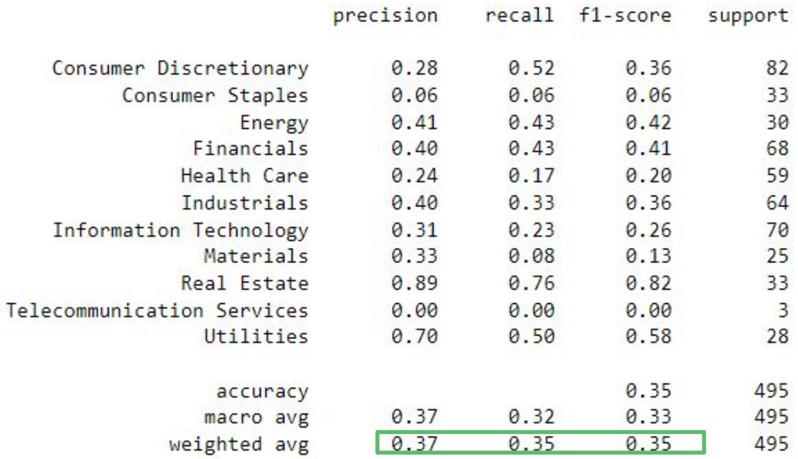
A) Benchmark sectors via the efficacy of a cross-validated K-NN model

B) Create new ‘sectors’, clusters, for each stock using Batch K Means

C) Compare the meaningful features of both categorization methods

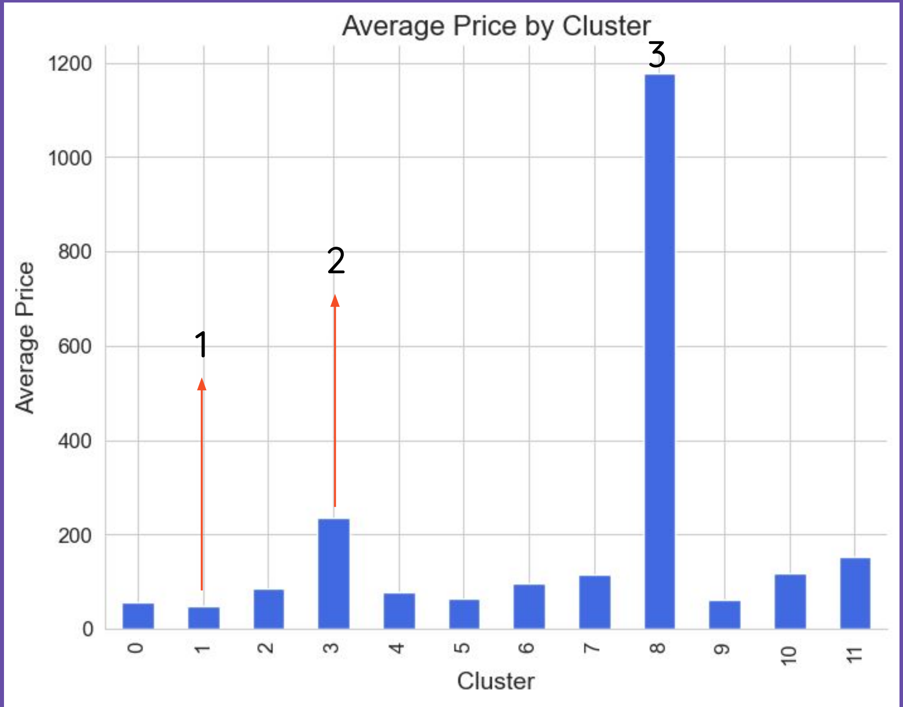


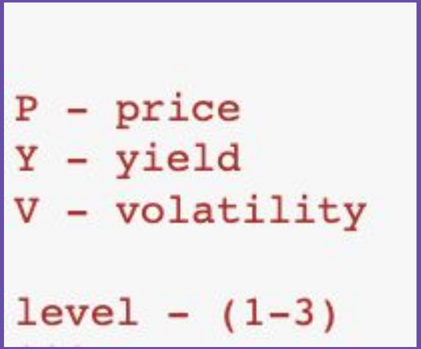
	Consumer Di	Consumer St	Energy	Financials	Health Care	Industrials	Information	Materials	Real Estate	Telecommuni	Utilities
Consumer Di	43	4	4	5	6	10	6	2	0	0	2
Consumer St	18	2	1	4	2	2	2	0	0	0	2
Energy	7	2	13	2	1	0	5	0	0	0	0
Financials	15	7	2	29	6	4	2	0	2	0	1
Health Care	18	4	1	6	10	6	13	1	0	0	0
Industrials	22	4	1	8	4	21	3	0	0	0	1
Information	16	9	4	11	9	4	16	1	0	0	0
Materials	9	2	3	1	2	4	2	2	0	0	0
Real Estate	2	0	0	4	1	0	1	0	25	0	0
Telecommuni	1	1	0	0	0	0	0	0	1	0	0
Utilities	4	1	3	3	1	1	1	0	0	0	14



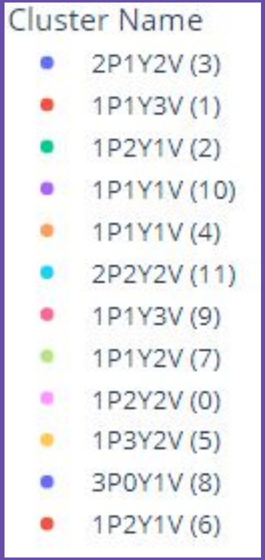
```
                            precision    recall  f1-score   support

    Consumer Discretionary       0.28      0.52      0.36        82
          Consumer Staples       0.06      0.06      0.06        33
                    Energy       0.41      0.43      0.42        30
                Financials       0.40      0.43      0.41        68
               Health Care       0.24      0.17      0.20        59
               Industrials       0.40      0.33      0.36        64
    Information Technology       0.31      0.23      0.26        70
                 Materials       0.33      0.08      0.13        25
               Real Estate       0.89      0.76      0.82        33
Telecommunication Services       0.00      0.00      0.00         3
                 Utilities       0.70      0.50      0.58        28

                  accuracy                           0.35       495
                 macro avg       0.37      0.32      0.33       495
              weighted avg       0.37      0.35      0.35       495
```

Cluster Creation




P - price
Y - yield
V - volatility

level - (1-3)


Cluster Name

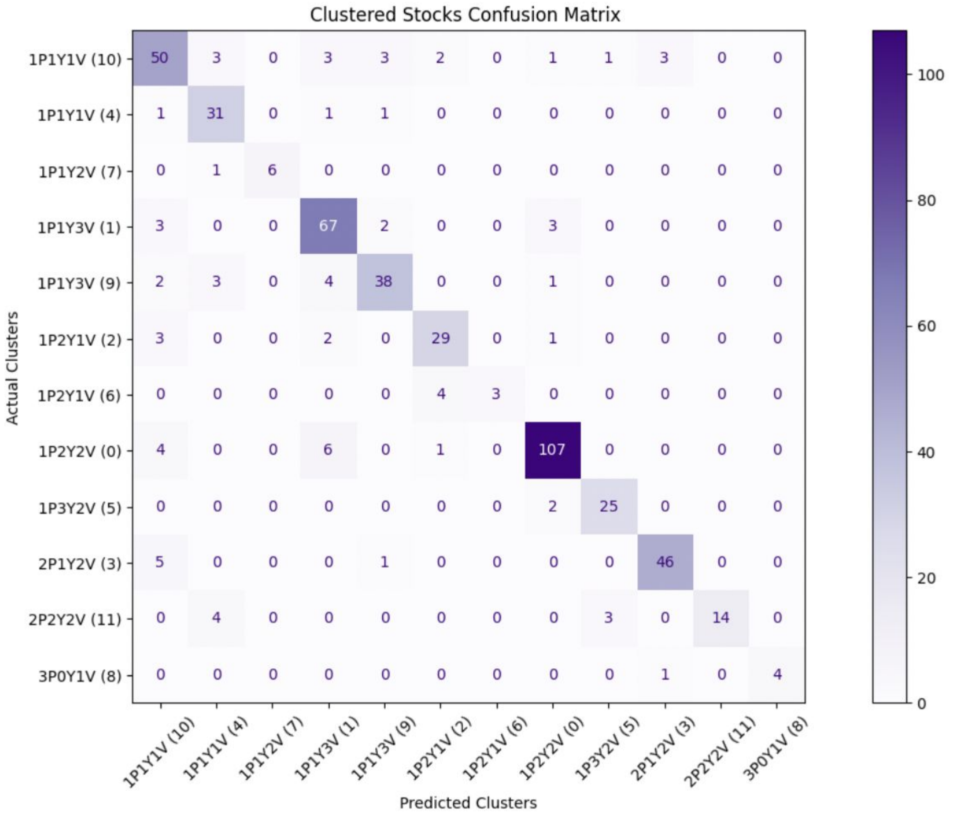
- ● 2P1Y2V (3)
- ● 1P1Y3V (1)
- ● 1P2Y1V (2)
- ● 1P1Y1V (10)
- ● 1P1Y1V (4)
- ● 2P2Y2V (11)
- ● 1P1Y3V (9)
- ● 1P1Y2V (7)
- ● 1P2Y2V (0)
- ● 1P3Y2V (5)
- ● 3P0Y1V (8)
- ● 1P2Y1V (6)

*For all visualizations/tests data is normalized s.t. col mean = 0 and std = 1.

Results

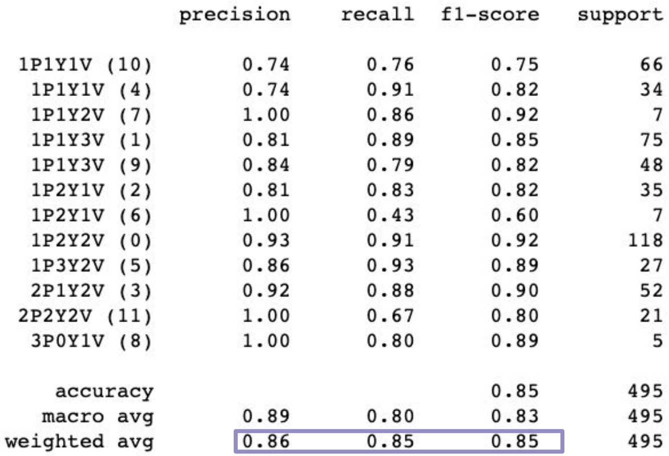
- Weighted average precisionis 86%

Fail or Success?



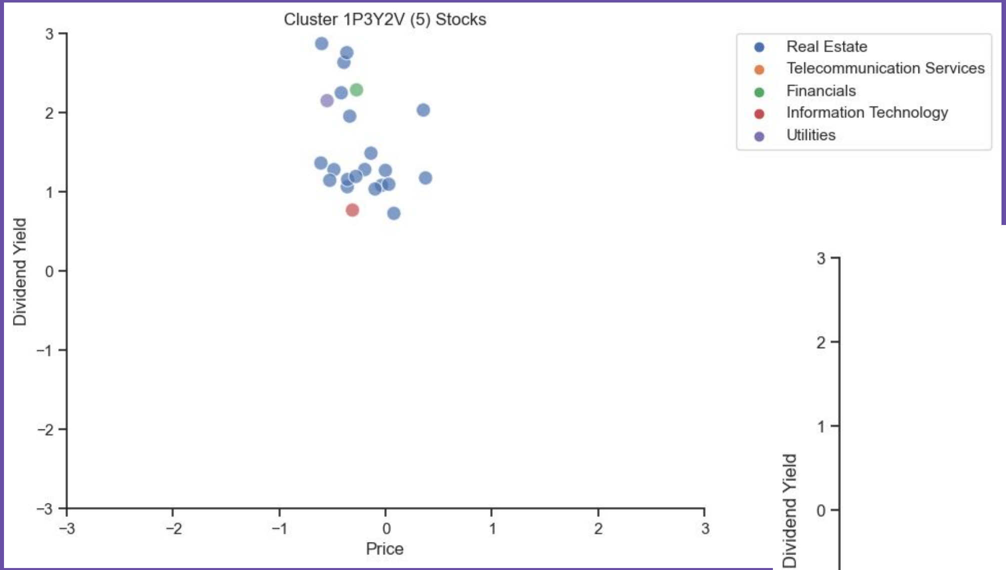
Clustered Stocks Confusion Matrix

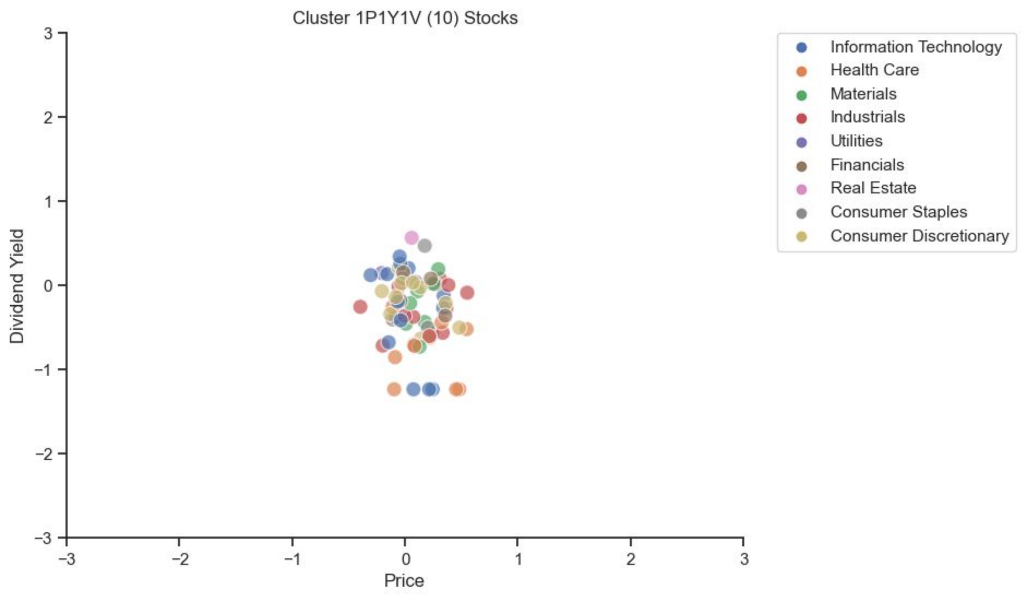
	1P1Y1V (10)	1P1Y1V (4)	1P1Y2V (7)	1P1Y3V (1)	1P1Y3V (9)	1P2Y1V (2)	1P2Y1V (6)	1P2Y2V (0)	1P3Y2V (5)	2P1Y2V (3)	2P2Y2V (11)	3P0Y1V (8)
1P1Y1V (10)	50	3	0	3	3	2	0	1	1	3	0	0
1P1Y1V (4)	1	31	0	1	1	0	0	0	0	0	0	0
1P1Y2V (7)	0	1	6	0	0	0	0	0	0	0	0	0
1P1Y3V (1)	3	0	0	67	2	0	0	3	0	0	0	0
1P1Y3V (9)	2	3	0	4	38	0	0	1	0	0	0	0
1P2Y1V (2)	3	0	0	2	0	29	0	1	0	0	0	0
1P2Y1V (6)	0	0	0	0	0	4	3	0	0	0	0	0
1P2Y2V (0)	4	0	0	6	0	1	0	107	0	0	0	0
1P3Y2V (5)	0	0	0	0	0	0	0	2	25	0	0	0
2P1Y2V (3)	5	0	0	0	1	0	0	0	0	46	0	0
2P2Y2V (11)	0	4	0	0	0	0	0	0	3	0	14	0
3P0Y1V (8)	0	0	0	0	0	0	0	0	0	1	0	4



```
                precision    recall  f1-score   support

 1P1Y1V (10)        0.74      0.76      0.75        66
  1P1Y1V (4)        0.74      0.91      0.82        34
  1P1Y2V (7)        1.00      0.86      0.92         7
  1P1Y3V (1)        0.81      0.89      0.85        75
  1P1Y3V (9)        0.84      0.79      0.82        48
  1P2Y1V (2)        0.81      0.83      0.82        35
  1P2Y1V (6)        1.00      0.43      0.60         7
  1P2Y2V (0)        0.93      0.91      0.92       118
  1P3Y2V (5)        0.86      0.93      0.89        27
  2P1Y2V (3)        0.92      0.88      0.90        52
 2P2Y2V (11)        1.00      0.67      0.80        21
  3P0Y1V (8)        1.00      0.80      0.89         5

    accuracy                            0.85       495
   macro avg        0.89      0.80      0.83       495
weighted avg        0.86      0.85      0.85       495
```

Sector Visuals





Final Thoughts

- Industry is not the best indication for how well a stock could do (as shown by our confusion matrices)
- Current methodology is subject to what is deemed as key ratios and metrics - may be better to fit the classifier on daily stock fluctuations

Thanks!

Any questions?