Stock Analysis

Jan D. Busse

February 19, 2018

Abstract

Companies, or rather their stocks, get grouped by industries. A widely used industry taxonomy is the Global Industry Classification Standard (GICS)[1]. This report examines if the classifications proposed by the GICS can be found while applying machine learning algorithms to the stock prize data. Therefore stocks from the S&P 500 in 2017 are analyzed and the results are compared to the GICS classification of the same stocks.

1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

2 Data Preparation

Before applying machine learning, preparing the data is crucial. In this case all stocks, that didn't offer the prize history for the whole year 2017, were discarded in the first place. For the remaining 500 stocks the closing prize for each day was chosen to perform all of the following analyses.

[1] https://www.msci.com/gics

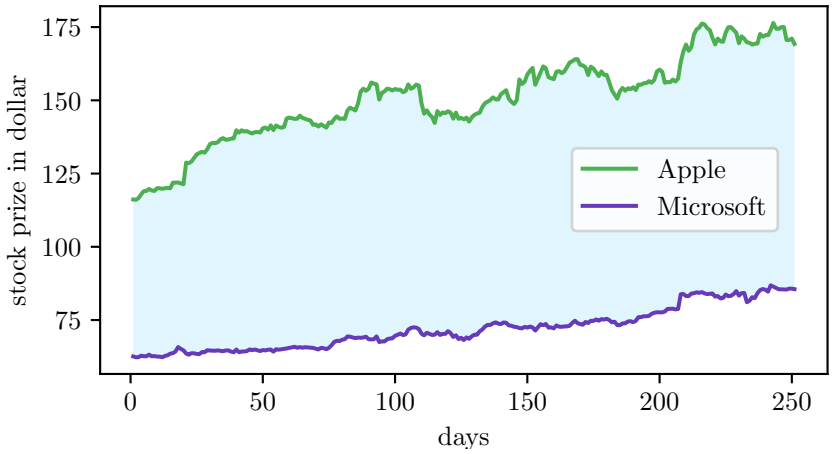

Figure 1: Stock prizes of Apple and Microsoft during 2017.

3 Results

3.1 k-means

Why k-means? k-means is a good spot to start when clustering. It is fast and one of the major downfalls, the number of clusters must be known in advance, is no problem in this case. Since we are looking for the GICS sectors, we assume the number of clusters is the same as GICS proposes, which are eleven major sectors[2].

For the first run k-means was applied directly to the untouched stock prizes. That means the distance between two stocks is defined by

$$d_{i,j} = \sum_{k=1}^{n} (p_{i,k} - p_{j,k})^2, \tag{1}$$

where $p_{i,k}$ is the closing prize of stock i on day k and n the total number of days in the dataset. In simpler terms, the distance of two stocks is defined by the squared area in between both charts. Figure 3.1 shows the distance for the stocks of Apple and Microsoft.

Table 1 pictures the crosstab between the GICS classification on the left and the clusters determined by the k-means algorithm at the top. A chi-squared contingency test results in a χ^2 value of 113.58, which corresponds to a p-value of 0.1667. Thus the Null hypothesis, that the GICS classification and k-means cluster are stochastically independent, cannot be

Table 1: Crosstab k-means clustering directly on the stock prize, ($\chi^2 = 113.58$, $p = 0.1667$).

	k-means cluster center										
GICS	0	1	2	3	4	5	6	7	8	9	10
10	11	0	3	0	0	0	11	1	0	0	6
15	5	0	8	0	1	1	5	1	0	0	4
20	6	0	11	0	7	0	22	7	0	3	11
25	24	1	7	1	3	2	24	5	1	0	15
30	4	0	6	0	0	0	12	2	0	0	10
35	5	0	9	0	4	1	12	11	1	2	14
40	14	0	7	0	2	1	20	6	0	0	17
45	17	2	11	0	2	0	11	7	0	0	19
50	2	0	0	0	0	0	1	0	0	0	0
55	8	0	3	0	0	0	9	0	0	0	8
60	9	0	8	0	2	1	7	2	0	0	4

discarded on a 5 % level of significance. This means that there is no significant correspondence between the GICS classes and the clusters found by k-means.

A possible cause for the lack of equivalence could be the way the distance is calculated. Equation 1 uses the absolute stock prizes. That way the distance of two stocks with a similar trend can be bigger, due to large differences in the absolute stock prize, than between two stocks with different development but close stock prizes. Imagine the Apple stock from figure 3.1 would be flipped left to right and then compare it to the original trend. Even tough the trends of these stocks is the exact opposite, the area between them will be less than the blue area in figure 3.1.

To further investigate that assumption it's necessary to take offset and amplitude into consideration. Offset refers to different stock prizes in the first place. The stock price for Apple starts at \$116, the one of Microsoft at \$63. But we are not interested in the absolute difference rather if both stocks develop the same way from there. Adjusting the stock price by the it's mean over the whole year of 2017 will remove the offset. But different amplitudes remain. Imagine both stocks would rise by ten percent. Then Apple's stock would be at \$128 and Microsoft's at \$69 which results in a difference of \$59. Before the difference was \$53, so the difference increased even tough both stocks grew at the same scale. To scale the amplitude we

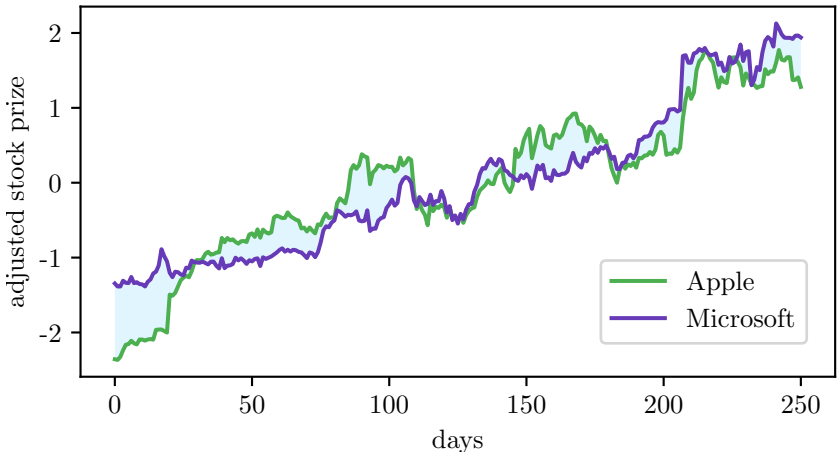

Figure 2: Adjusted stock prizes of Apple and Microsoft during 2017.

will divide the offset adjusted stock prices by the standard deviation over the whole year. The adjusted price is calculated by

$$p_{adj,k} = \frac{p_k - \overline{p}}{\sigma_p} \tag{2}$$

where p_k is the closing prize of the stock on day k, $\overline{p}$ the mean of all stock prizes for that stock and σ_p the standard deviation.

After the adjustment the k-means algorithm is run again. It's results are shown in table 2. The contingency test concludes in a χ^2 value of 274.80, corresponding to a p-value of less than 0.001. A G-Test results shows a similar test statistic. Therefore in this case the cluster centers found by the k-means algorithm are highly correlated to the sectors defined by GICS.

Even tough the test statistics appear to be very well, looking at the crosstab in table 2 clearly shows, that there's no clear mapping between the cluster centers and the GICS sectors. The ideal result obviously would be one cluster per sector, but all clusters are actually scattered across the different sectors.

To test this we can take the so called Rand index into consideration [3]. Which measures the similarity of two data clusterings. In this case we use the Adjusted Rand Index (ADI) [1], that yields in a value between -1 and 1. In case the ADI is close to zero both clusterings do not agree in any way, whereas a value close to one states that both clusterings are similar.

The ADI for table 2 results in a value of 0.0368, which confirms our

Table 2: Crosstab k-means clustering with offset and amplitude scaling, ($\chi^2 = 274.80$, $p \ll 0.001$).

	k-means cluster center										
GICS	0	1	2	3	4	5	6	7	8	9	10
10	9	0	0	3	4	1	1	9	0	3	2
15	1	4	2	10	2	0	0	1	1	3	1
20	2	8	3	24	4	0	5	2	7	12	0
25	8	6	5	17	6	5	13	6	3	10	4
30	4	2	5	4	1	2	6	2	4	3	1
35	2	13	6	16	0	5	6	0	3	5	3
40	0	3	0	21	15	4	1	1	3	18	1
45	4	12	2	34	2	1	0	1	3	6	4
50	0	0	0	0	1	0	1	1	0	0	0
55	0	10	0	10	0	3	2	0	1	2	0
60	4	8	0	5	0	3	1	7	1	3	1

observation.

References

[1] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[2] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[3] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.