Stock Analysis

Jan D. Busse

February 15, 2018

Abstract

Companies, or rather their stocks, get grouped by industries. A widely used industry taxonomy is the Global Industry Classification Standard (GICS). 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 then are opposed to the GICS classification of the same stocks.

1 Introduction

Introduction blabla

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.

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

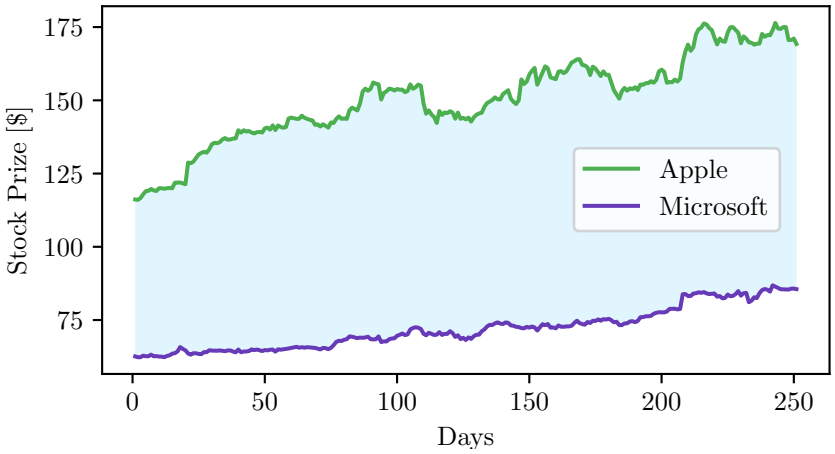

Figure 1: Stock prizes of Apple and Microsoft during 2017

sectors, we assume the number of clusters is the same as GICS proposes, which are eleven major sectors.

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 of each other cannot be 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

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

	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

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 compare the trends of the two stocks to each other. Therefore the