THE BUCHAREST UNIVERSITY OF ECONOMIC STUDIES

THE FACULTY OF CYBERNETICS, STATISTICS AND ECONOMIC INFORMATICS

ECONOMIC INFORMATICS SPECIALIZATION

Comparing stocks performances through an unsupervised classification

COORDINATOR:

Conf. Univ. Dr. Claudiu **VINȚE**

STUDENT:

Dincă Laurențiu – Gabriel

Group 1098, G Series

BUCHAREST

2022

1. **Description of the chosen topic and selected data**

The main purpose of the current analysis which covers the automate unsupervised learning is to group different American companies, listed on the stock exchange, according to their financial performance. In this regard, 31 companies were selected and 10 financial indicators as it follows:

* **Income growth** (annual percentage increase);
* **Operating income:** The income obtained from the company’s economic activity after deductions(i. e. salaries of the employees) and depreciation;
* **Operating Expenses**: Resulted expenses from the company’s economic activity;
* **Total Assets**;
* **Total Debt**;
* **Operating Cash Flow**: The total amount of liquidities held by the company. Usually, the cash flow is assessed in relation to the total debts and may represent a measure of the company’s solvency;
* **Investing Cash Flow** : The total amount of money which have been invested in assets that will generate non-operational profit in the future;
* **Operational profitability** (EBITDA Marginal): Represents the computed indicator computed as a ratio between EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization) and total sales;
* **(Return on Assets – ROA):** The indicator is computed as the ratio of net profit to the total assets of the company. This index reflects the efficiency with which assets owned by a company are used;
* **(Return on Equities – ROE):** The indicator is computed as the ratio between the net profit obtained by the company and the equity, the latter representing practically the shareholders’ contribution to the financing of the business.

Is there any possibility to create a model that might be able to make a difference between stocks that perform well on the market and those who do not? Therefore, how can we leverage this knowledge in order to predict which stock will be worth buying in the future? Despite that, would it be possible for us to achieve such a thing based on some financial indicators already provided? In order to answer, at least for a general matter to this question we are going to analyze how the above-mentioned indicators might interfere within a company and its financial efficiency.

Another important aspect which worth to be mentioned is the nature of the dataset used for the analysis, built at the end of 2018 as a good predictor for the stock performance in 2019, dataset available on the Kaggle, the global online community of data scientists and machine learning practitioners.

Additional information is also available here: [Beat US Stock market (2019 edition) | Kaggle](https://www.kaggle.com/cnic92/beat-us-stock-market-data)

1. **Unsupervised classification**

Clustering analysis aims to group shapes into multiple classes according to certain common characteristic features. Therefore, it is desired that the observations included in a given class should be as similar as possible, while the observations included in different classes should be as little similar as possible. In this regard, we are going to construct the matrix of distances, in order to assess the de(similarity) between observations.

The reason for choosing this type of data analysis consists in a deeper understanding of how a cluster is created, in particular how several companies may be placed on the same level of efficiency from an economic point of view, even though, at a first look, the “accounting numbers” seems to be quite different.

 **Euclidian distance**

The Euclidian distance between two vectors(two observations if we follow the scenario above) is computed by using the following formula:

where xi represents the values of the first observations for the “i” indicator, yi represents the values of the second observation for the “i” indicator, and m represents the total number of indicators used.

In order to better understand how an unsupervised classification represents an explanatory method in which the employed data are values which highlight the relationship among the included companies we will try to exemplify three different types of hierarchical clustering.

* 1. **Centroid method**

Chart, histogram

Description automatically generatedWe are going to apply the Centroid method in order to form the corresponding clusters.

Figure 1 - Hierarchical Classification(Centroid-Euclidian)

Using the dendrogram above, we are going to determine the optimum number of classes, looking through the largest distance between two consecutive steps. Furthermore, we will draw a horizontal line which is going to intersect the hierarchical structure in three points. To conclude, according to this criteria, the optimum number of classes is equal to 3.

The centroid represents the mean of an indicator among the given dataset, by considering only the observations contained by that particular cluster. By comparing the values of the centroids, we might be able to assert the fact that in the first class we associated the companies more developed from a financial point of view, while in the second class there will be a number of three companies less developed in terms of financial indices.

Another aspect worth mentioning is the value of the threshold, which, for this method is equal to 3.803.

* 1. **Ward Method**

Next, we are going to apply the Ward Method of ascending hierarchy in order to form the corresponding clusters. The results might be observed below:**Chart

Description automatically generated**

Figure - Hierarchical Classification(Ward-Euclidian)

By using the dendrogram above, we determined the optimum number of classes which is similar with the centroid method. Therefore, we will draw a line which will intersect the hierarchical structure in three points, where the threshold has reached the value of 6.929.

Also, we might see that the first cluster contains a quite unexpected number of 11 companies, the second one contains three companies, while the last one comprehends all of the remaining ones.

2.2 **Complete Method**

Complete or maximum aggregation method defines cluster proximity as the furthest distance between two points :

Chart

Description automatically generated

Figure 3 - Hierarchical Classification(Centroid-Euclidian)

Using the dendrogram above, we are going to determine the optimum number of classes, looking through the largest distance between two consecutive steps. Furthermore, we will draw a horizontal line which is going to intersect the hierarchical structure in two points, where the value of the threshold is equal to 7.578. To conclude, according to this criteria, the optimum number of classes is equal to 2.

Since the centroid, respectively Ward values are not quite similar to those obtained when applying the maximum method, the two classes will be now retaining their economic interpretation, instead of three. However, five companies will be considered medium developed in terms of financial performance.

**2.3 Hierarchical Classification of Variables**

**Chart, histogram, box and whisker chart

Description automatically generated**As a last step, we are going to describe the importance of the indicators previously used by computing a hierarchical classification.

Figure 4 - Hierarchical Classification of Variables

It is interesting to see how, based on the dendrogram above, the financial indicators are associated when it comes to creating clusters, and even more, that only 7 of them are actually correlated in respect to our analysis. Therefore, we are able to conclude that the most relevant indicators for predicting a company’s financial potential of performance are the ROE, ROA, Investing Cash Flow and Operating Profitability indices.

1. **Conclusions and Future work**

After running the corresponding data analysis methods and comparing the obtained results, we may come back to our hypothesis and affirm the fact that by choosing the right financial indicators, we are able to determine the future performance of a company on the stock market. In addition, I would like to continue my research on this topic by building my own statistical model, based on this experience.