Merge

Volume 3 Article 5

2019

Reliability of Clustering in Forecasting Stock Prices of Companies Traded on the Stock Exchanges

Shristi Dhakal

Reliability of Clustering in Forecasting Stock Prices of Companies Traded on the Stock Exchanges

Shristi Dhakal

Mississippi University for Women

Abstract

Whether stock prices can be accurately forecasted or not is usually associated with whether markets are efficient or not. The idea of market efficiency suggested by the Efficient Market Hypothesis has been debated among financial professionals for a long time, especially due to the occurrence of financial bubbles in the past. Some argue that stock prices prediction is no different than the results of "a series of tosses of a coin, rolls of a die, or spins of a roulette wheel," while others argue that stock prices are affected by past patterns, which can be used to forecast future prices [11]. Although a concrete answer has yet to be found on the behavior of the stock market, researchers have continued exploring the topic and have established various quantitative models for forecasting, one of which is clustering. This paper evaluates the application of the clustering method of stock forecasting by analyzing the financial statements of technology companies over a period of four years.

Keywords: stock prediction, clustering

1. Introduction

Stock market prediction is a complex topic. It requires analyzing a lot of past and present data. With markets fluctuating every now and then, the amount of information is only increasing day by day. So, there are more chances that predictions will fail if our understanding of the huge amount of available data is not aided by analyses using accurate mathematical methods. Among many prediction methods that are mathematical, clustering is one of them, and it is used to identify patterns and behaviors using correlations as measure of similarity. The application of clustering is interpreted differently by various studies depending on the data and outcomes of the study, but it is believed to have yielded consistent results in most cases when applied with stocks.

1.1 Background

1.1.1 Efficient Market Hypothesis

The Efficient Market Hypothesis (EMH) is an investment theory which states that stock markets are efficient and reflect all information about individual stocks and the stock market in general [2, 5]. The theory is related to the idea of a "random walk," which suggests that prices move randomly [5]. The reasoning behind the random walk idea is that if information about a company flows in the market without hindrances and it quickly gets reflected in the stock prices, then tomorrow's prices will reflect only tomorrow's news and will not depend on the prices changes today [5]. Since tomorrow's news is unpredictable, stock price movements must also be unpredictable and random [5]. Hence, prices fully reflect all known information, and therefore, investors, neither through technical analysis, which is the study of past price movements to predict future price movements, nor fundamental analysis, which is the analysis of financial

information such as asset values and profit margins, can obtain future information that helps to pick ‘winning’ stocks and achieve returns that are higher than average without taking comparable risk [5].

The rational expectations set by the EMH was believed by many around the 1970s until behavioral economists started questioning it due to the occurrence of anomalies that did not quite fit with the idea of the efficient markets theory [10]. Economists who analyzed the psychological and behavioral aspects of finance argued that the EMH did not explain market irrationalities such as the crash of 1987 and the internet bubble of the 1990s enough. Some of these economists, by studying the psychological elements of stock prices, even claimed that future prices are somewhat predictable based on the past patterns and some fundamental valuation metrics, which may allow investors to select stocks that can provide risk-adjusted returns [5].

These two opposite arguments have sparked debates on market efficiency and its predictability for many years. Today, the two components of the EMH- 1) stock prices quickly reflect all relevant information, and 2) if part 1 is true, investors cannot select ‘winning’ stocks and beat the market, are argued separately. Some experts agree with the first part, whereas some disagree, but even those who disagree with the first part would agree with the second part because investing in stocks that perform better than the market average appears to be difficult.

1.1.2 Forecasting in Stock Market

Regardless of the controversies in the stock market theories, forecasting is considered important by many in future decision making. Past patterns have been useful in detecting future behaviors. Investors such as Warren Buffet have forecasted prices that have eventually turned out to be true and have been outperforming the markets for many years. Although the success of

these investors is debated as being chance events by proponents of the EMH, it certainly does not appeal in the same way to the opponents, which is why complex mathematical models like clustering have been developed.

1.2 Cluster Analysis

Cluster analysis is the method of grouping objects into similar and dissimilar groups so that objects in one group are similar to one another and are different from the objects in other groups [6]. There are typically two methods of clustering: hierarchical and partitional clustering. This paper uses hierarchical clustering, which is based on the relationship between the two nearest clusters and, unlike other clustering, the data sets "are not partitioned into a particular cluster in a single step." Rather, they go through a series of partitions, from "a single cluster containing all objects to N clusters each containing a single object" [6].

2. Related Work

Patterns drawn from clustering have been used in the past to effectively forecast stock prices. Babu et al. (2012) have suggested that clustering can be helpful for investors in finding the interrelationship between the long term underlying trends of stock prices that is normally hidden by the short-term fluctuations. Lee et al. (2010) have analyzed the role of clustering in investigating the relationship between financial reports and short-term stock prices and have recommended better techniques for better predictions. Similarly, several other studies have suggested that clustering is valuable in time series forecasting and has produced good results.

3. Hypothesis

The objective of this paper is to see if clustering can be used to beat the market. So, based on the objective, below are the null and the alternative hypotheses:

Null Hypothesis: There is no difference in the amount of profit that can be earned using clustering.

Alternative Hypothesis: There is a difference in the amount of profit that can be earned using clustering.

4. Methodology

The primary method used for this project is data collection and analysis. Information from the financial statements are used to predict the profit margin.

4.1. Financial Statements

Financial statements contain information such as the assets, liabilities, equity, cash inflows and outflows, revenues, income/loss, and profits of a company. It comprises three parts: Balance Sheet, Income Statement, and Cash Flow Statement. The Balance Sheet provides an overview of a company's assets, liabilities, and shareholder's equity; the Income Statement gives an overview of revenues, expenses, net income, and earnings per share; and the Cash Flow Statement combines both and gives an overall picture of the operating, investing, and financing activities of a company. Such information is considered useful when determining the worth of a company for investment. Studies suggest that financial statements can identify fundamentals that are not visible in prices, and there exists a relation between stock prices and earnings of a company, which could be used for making informed investments decisions. [6].

The information from financial statements can be interpreted differently as it can infer multiple things. In this paper, the information is used as a similarity measure to group companies into clusters. Grouping of companies based on financial performance gives investors a general idea of which companies perform better financially, and this could be a valuable information while selecting companies to invest in.

5. Discussion of data

In a perfect world, 60 companies are required to get sufficient results so that two groups can be formed with 30 companies in each group. Also, it is preferred that the companies be in the same industry to make relevant comparisons. Hence, 60 top performing technology companies traded on the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotations (NASDAQ) were randomly selected and following information from their financial statements from 2014 to 2017 was recorded on a spreadsheet:

i. From balance sheet: cash, current assets, fixed assets, current liabilities, long term debt, and equity.

ii. From income statement: Earnings Before Interest and Taxes (EBIT) and profit margin.

Large corporations have their financial statements publicly available in various websites. Because it is simple and easier to navigate, Google Finance was used to gather the data for this study.

Figure 1: List of companies for clustering

Companies	Symbol
Apple Inc	AAPL
Alphabet Inc	GOOGL
Microsoft Corporation	MSFT
Facebook	FB
Intel Corporation	INTC
Oracle Corporation	ORCL
Cisco Systems, Inc.	CSCO
International Business Machines Corporation	IBM
SAP SE	SAP
NVIDIA Corporation	NVDA
QUALCOMM Incorporated	QCOM
Texas Instruments Incorporated	TXN
NTT DOCOMO, Inc	DCM
Adobe Systems Incorporated	ADBE
Salesforce.com	CRM
Baidu, Inc.	BIDU
Applied Materials, Inc.	AMAT
Illinois Tool Works, Inc.	ITW
Micron Technology, Inc.	MU
Automatic Data Processing, Inc.	ADP
Activision Blizzard, Inc	ATVI
Cognizant Technology Solutions Corporation	CTSH
Intuit Inc.	INTU
NXP Semiconductors N.V.	NXPI
HP Inc.	HPQ
Electronic Arts Inc.	EA
Infosys Limited	INFY
Lam Research Corporation	LRCX
Analog Devices, Inc.	ADI
Nokia Corporation	NOK
Autodesk, Inc.	ADSK
Fiserv, Inc.	FISV
Western Digital Corporation	WDC
Kyocera Corporation	KYO
Wipro Limited	WIT
Workday, Inc.	WDAY
Red Hat, Inc.	RHT
Hewlett Packard Enterprise Company	HPE
Cerner Corporation	CERN
ServiceNow, Inc.	NOW
Microchip Technology Incorporated	MCHP
STMicroelectronics N.V.	STM
Skyworks Solutions, Inc.	SWKS

Maxim Integrated Products, Inc.	MXIM
Motorola Solutions, Inc.	MSI
Check Point Software Tech Ltd.	CHKP
Symantec Corporation	SYMC
Xilinx, Inc.	XLNX
L3 Technologies Inc	LLL
Dover Corp	DOV
NetApp Inc	NTAP
Palo Alto Networks Inc	PANW
CA, Inc.	CA
Seagate Technology PLC	STX
United Rentals, Inc.	URI
ASML Holding NV (ADR)	ASML
Citrix Systems, Inc.	CTXS
Yandex N.V.	YNDX
Synopsys, Inc.	SNPS
Splunk Inc.	SPLK

5.1. Common Size Financial Statements, Financial Ratios, and Time Series Data

Common size financial statements are ideal when making comparisons between companies of different sizes [9]. They provide a picture of how a company is doing in terms of total assets and total revenues, and what the numbers compare to other companies in the same industry or in different industries of different sizes. In other words, the "common-size" statements can be used for both "intercompany" and "intracompany" comparison [9]. The numbers in the financial statements are converted to percentages so that there is a common base [8]. The total of assets or liabilities and capital are assumed to be 100%, and rest of the financial items such as cash, current assets, and current liabilities are then calculated as percentage of total assets and/or total liabilities or equity. Similarly, in the income statement, total revenues/sales are assumed to be 100%, and items such as profit margin and Earnings Before Interest and Taxes are computed in relation to the total revenue. For instance, Facebook, one of the companies used in this study, had total cash of 2.15 billion, total assets of 39.96 billion, current liabilities of 1.42

billion, and total liabilities of 3.87 billion in 2014. When converted to percentages, this means that cash was 5.39% of total assets and current liabilities was 36.80% of total liabilities, which suggests that Facebook had significantly higher percentage of current liabilities than percentage of cash in 2014. Similar comparisons can be made between two or more companies by comparing the percentages of financial items. For this study, cash, current assets, fixed assets, current liabilities, long-term debt, equity, EBIT, and profit margin were converted to their respective percentages and a common size financial statement of the 60 companies was created.

The amounts that are converted into percentages can also be called financial ratios. Financial ratios give "the numeric outcome obtained by dividing one financial data with other and is used to express the relativity of different financial variables" [9]. The difference between common size statements and financial ratios is that the calculation of financial ratios uses data from one, two or more financial statements, whereas the common size statement typically uses data from the same financial statement for which it is being computed. In other words, a common size balance sheet will have items from the balance sheet only and a common size income statement will have items from the income statement only, unlike in the case of a financial ratio such as total assets turnover ratio, which is computed by dividing total sales by total assets, using information from both the income statement and the balance sheet. In this study, there are no ratios calculated from multiple financial statements, so the terms- common size statements and financial ratios may be used interchangeably.

The collected data includes a time horizon of four years, so there are some patterns that provide additional information about the performances of the selected companies over the years. Analysis done by looking at the data from several years or quarters is known as time series analysis [9]. Time series technique can be used for forecasting future values or prices, in which it

utilizes current data to make accurate predictions about future unknown data values [3]. The information obtained through this technique can be valuable to investors estimating their future returns.

In this paper, financial ratios, common size statements, and time series data serve as a basis to group the companies into clusters. Statistical tools are utilized to explain the behavior of the ratios and the compiled financial data.

Figure 2: Sample of the Common size statement (2014-2017)

		Cash in %	CA in %	FA in %	CL in %	LTD in %	Equity in %	EBIT in %	Profit Margin in %
Facebook	2014	5.39	33.50	14.47	36.80	3.07	90.32	40.06	82.73
	2015	6.48	43.82	15.83	37.10	2.06	89.50	34.72	84.01
	2016	11.61	52.96	18.17	49.85	0.00	91.12	44.96	86.29
	2017	6.94	57.45	21.69	36.95	0.00	87.96	49.70	86.58
Oracle	2014	19.69	53.32	6.91	33.16	49.49	51.93	38.56	81.09
	2015	19.58	56.97	6.58	24.57	62.63	43.88	36.29	80.30
	2016	17.96	57.33	7.22	26.52	60.28	42.15	34.02	79.81
	2017	16.14	55.20	7.63	29.80	58.09	39.90	33.69	80.20
HP Inc.	2014	14.66	48.59	25.44	57.19	20.97	25.90	7.51	19.81
	2015	7.10	48.45	5.36	53.33	8.44	25.98	7.62	19.31
	2016	21.69	63.71	20.99	57.21	20.49	-13.42	7.36	18.65
	2017	21.26	67.81	18.10	61.71	18.58	10.35	6.76	18.40

6. Data Analysis

Because the clustering software was not able to comprehend partially available data, 13 companies that had one or more unavailable amounts for financial items were taken off the list. Hence, only 47 companies out of the 60 contributed to the results.

6.1 Clusters

With the financial data from 2014 to 2016, two big clusters were formed; one with 41 companies (Group 1) and another with only 6 companies (Group 2). These groups are very

uneven as there is a big difference in the number of companies present in the two groups. The 6 companies are very different from the companies present in the other group.

Figure 3: Dendrogram of the clusters (1st dataset)

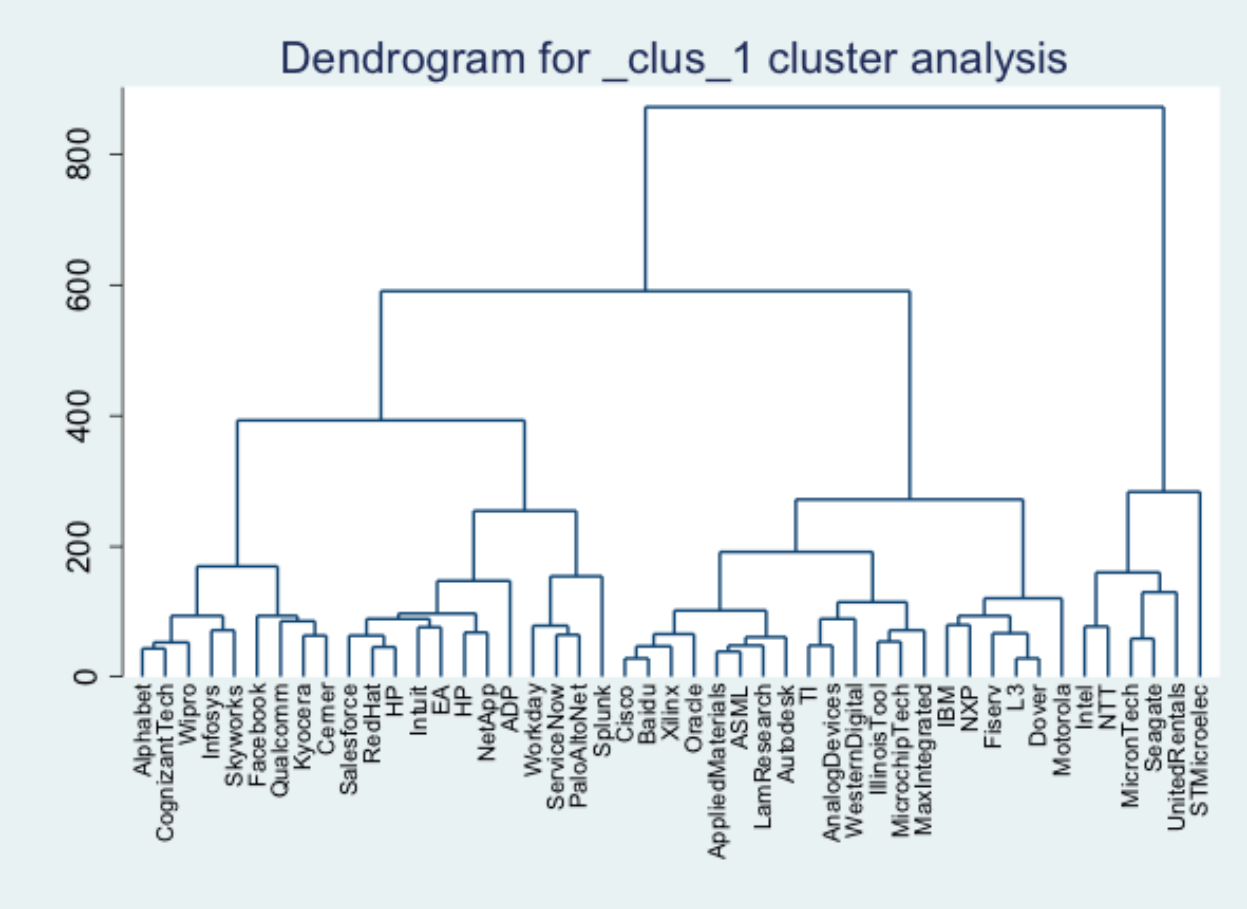


Figure 3 shows the formation of clusters of 47 companies of which financial ratios were entered in the software. As can be seen in the figure, Intuit and Electronic Arts (EA) are in one cluster, which means that they share more similar characteristics with each other than with IBM, Oracle, Intel, or any other company. Same applies to other companies in the clusters. And, although there are several clusters formed, there are two big clusters that look very distinct than the other clusters. A t-test is run between these two clusters and following results are obtained:

Variable: Cash

Cluster	Mean
1	18.91372
2	13.28407
Total	18.19504

t = 0.8577

Pr (|T|>|t| = 0.3956

Variable: Current Assets

Cluster	Mean
1	54.60374
2	39.50438
Total	52.67616

t = 2.0983

Pr (|T|>|t| = 0.0415

Variable: Fixed Assets

Cluster	Mean
1	24.18187
2	108.8208
Total	34.98684

t = -9.2135

Pr (|T|>|t| = 0.0000

Variable: Current Liabilities

Cluster	Mean
1	51.52622
2	40.74749
Total	50.15021

t = 1.2518

Pr (|T|>|t| = 0.2171

Variable: Long-term debt

Cluster	Mean
1	28.93337
2	42.14025
Total	30.61935

t = -1.4883

Pr (|T|>|t| = 0.1436

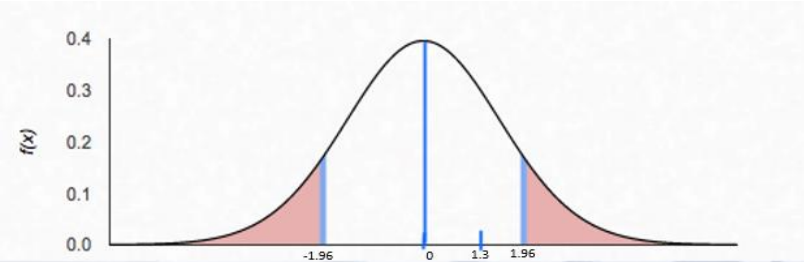
Variable: Equity

Cluster	Mean
1	52.82331
2	47.32542
Total	52.12146

t = 0.6167

Pr (|T|>|t| = 0.5405

Figure 4: T-distribution curve



Companies in Group 1 have slightly more cash, current liabilities, and equity, slightly less long-term debt but significantly more current assets and significantly less fixed assets than companies in group 2. Having more cash and current assets generally means that a company is more liquid and can recover quickly in case of bankruptcy. More current liabilities and more equity infer that a company has more short-term obligations and more ownership of assets respectively. Less fixed assets indicate that a company might be under-invested in plant, property, and equipment, whereas less long-term debt indicates that a company may be paying less interests, which can be good as the money can be allocated to other important areas of business.

The summary of the results in the balance sheet below provides overall view of the financial items in the two groups:

Assets	**Liabilities + Owner's Equity**
Current Assets (+)	Current Liabilities
Fixed Assets (-)	Long-term Debt
	Equity

**The items with statistically significant differences is marked with a plus or a negative sign.*

As seen in the above table, the significant differences are only on the assets side. Among the two groups, the difference in the amount of cash, current liabilities, and equity are not statistically significant as the t-values fall within the limits. With 95% confidence interval, for the variables to be statistically significant, their t-values should not fall under the limits of the curve, i.e. the values should be either lower than -1.96 or higher than 1.96 (represented by the shaded area in Figure 4). Cash, C.L., and equity have t-values of 0.86, 1.25, and -1.49 respectively, hence have no statistically significant differences in the average amounts. Current assets and fixed assets, however, have respective t-values of 2.10 and -9.21, both of which fall

beyond the limits of the curve, thus resulting in statistically significant differences in the average amounts of the financial items.

In summary, the clustering categorized the companies into two distinct groups and the companies in these groups vary by the use of their assets. Next, we see if this difference in assets leads to the differences in profit margins.

6.2 Profit Margin Prediction

First Dataset:

Financial ratios from 2014 to 2016 were used to categorize the companies into clusters to predict the profit margin for 2017. T-test was run and following information on profitability was obtained for the two groups:

Variable: Profit Margin

Cluster	Mean
1	55.34264
2	44.61156
Total	53.97272

t = 1.2974

Pr (|T|>|t| = 0.2011

The above table shows that Group 1 has higher profit margin on average than group 2. If there is no significant difference in the profit margin, there is a 95% chance that the t-score will be between 1.96 and -1.96. With the computed t-score of 1.2974, our results show no statistically significant difference in the amount of profit that can be earned. So, if investors were to invest in

these companies, their returns would likely not vary by exceedingly high amounts between the two groups. In other words, the average return being slightly higher in group 1 does not make a huge impact statistically.

Second Dataset:

To see if any major difference can be found with the alteration of available data, another t-test was run. This time, financial ratios from all four years: 2014, 2015, 2016, and 2017 were considered. A more balanced group was formed, with 20 companies in one group and 17 in the other group.

The results, however, were not very different than that of the first dataset.

Variable: Profit Margin

Cluster	Mean
1	60.23522
2	51.3737
Total	53.97272

t = 1.4101

Pr (|T| > |t|) = 0.1673

The t-value is slightly higher than that of the results of the first dataset, but it still lies within the limits in the curve. The probability of beating the average profit margin is only about 17%. Hence, no statistically significant difference in profitability between the clusters was found with the additional data from 2017.

Third Dataset:

A Duda-Hart stopping rule was run to find the optimal number of clusters, that maximizes between groups differences and minimizes within group differences. The results are summarized below:

clus	Obs	Rank Sum
1	9	195.00
2	12	380.00
3	20	455.00
4	5	87.00
5	1	11.00

**Obs = No. of companies*

Chi-squared = 6.236, d.f. = 4

Probability = 0.1822

The program categorized the companies into 5 clusters this time, which seems to fit the data well because the groups are more evenly divided. However, even with the more balanced clusters, the probability of earning a profit statistically higher is still very low (with a Chi-square of 6.24 and p-value of 0.18).

7. Conclusions & Future Work

The results of this study show no statistically significant difference in the profit margin between the groups, which reinforces the idea of the Efficient Market Hypothesis. It appears to be difficult to earn significantly higher amount of profit than the market average. But even though the difference in the profitability is not statistically significant, the economic significance can be huge. Referring to the first and the second dataset, an investor may not be able to earn a

statistically significant amount of money, but he/she may still be able to earn more if invested in Group 1 companies than in the Group 2 companies because the average profit margin is slightly higher in Group 1 in both cases. Even if it is just $500 or $1000 that an individual investor may earn, that could contribute to a huge amount economically when the earnings of all investors are pooled together. That is something to research more in the future.

Also, to derive more accurate conclusions about the predictability of the stock prices, future work can be done in testing the correlations between the profit margin and the stock prices and/or calculating the price-to-earnings ratio. Researchers can also include a larger number of companies and more time horizon for more data. Because the clustering software was not able to include companies that had incomplete data in this paper, it left us with 13 less companies than what we initially had for results. Whether including more companies and/or increasing the length of time changes the results or not, it would still be interesting to see what type of clusters can be formed. Furthermore, researchers can apply clustering to predict prices for other industries that are not as affected by speculation as technology and compare the results.

References:

1. Babu, S., Geethanjali, N., & Satyanarayana, B. (2012). Clustering Approach to Stock Market Prediction. *Int. J. Advanced Networking and Applications, 03*(04), 1290. Retrieved from http://www.ijana.in/papers/V3I4-10.pdf

2. Fama, E.F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance, 25*(2), 383-417. Retrieved from http://efinance.org.cn/cn/fm/Efficient%20Capital%20Markets%20A%20Review%20of%20Theory%20and%20Empirical%20Work.pdf

3. Kedia, V., Thummala, V. & Karlapalem, K. (2005). Time Series Forecasting through Clustering – A Case Study. *Duke University Department of Computer Science,* 183- 191. Retrieved from https://users.cs.duke.edu/~vamsi/papers/clustering.pdf

4. Lee, Anthony J.T., Lin, Ming- Chih, Kao, Rung- Tai, and Chen, Kuo- Tay. An Effective Clustering Approach to Stock Market Prediction. *PACIS 2010 Proceedings*, 54. Retrieved from https://pdfs.semanticscholar.org/9c31/6ac1192c4c76127c4916af29da21129b923c.pdf

5. Malkiel, B.G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives, 17*(1), 59-82. Retrieved from https://pdfs.semanticscholar.org/1e96/fb60c364c5ce5e21964c0ac992afb3507b75.pdf

6. Marinova-Boncheva, V. & Shelestov, A. Y. (2008). Using the agglomerative method of hierarchical clustering as a data mining tool in capital market. *International Journal "Information Theories & Application", 15*, 382-386. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.601.9161&rep=rep1&type=pdf

7. Razdar, M. R., & Ansari, M. (2013). A Study of Stock Price and Profitability Ratios in Tehran Stock Exchange (TSE). *2nd International Conference on Emerging Trends in Finance and Accounting.* Retrieved March 20, 2018, from https://www.researchgate.net/publication/287444828_A_STUDY_OF_STOCK_PRICE_AND_PROFITABILITY_RATIOS_IN_TEHRAN_STOCK_EXCHANGE_TSE

8. S., S. (2016). An Empirical Study on Balance Sheet Analysis. *Proceedings of the Fifth European Academic Research Conference on Global Business, Economics, Finance and Banking (EAR16Turkey Conference).* Retrieved March 23, 2018, from http://www.globalbizresearch.org/Turky_Conference_2016_Dec/docs/doc/2. Finance Accounting & Banking/I666.pdf

9. Sharma, R. (2012). Comparing and Analyzing Financial Statements to Make an Investment Decision: Case Study of Automotive Industry. Retrieved March 23, 2018, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.840.7999&rep=rep1&type=pdf

10. Shiller, R.J. (2003). From Efficient Markets Theory to Behavioral Finance. *Journal of Economic Perspectives, 17*(1), 83-104. Retrieved from http://www.econ.yale.edu/~shiller/pubs/p1055.pdf

11. Tryfos, P. (2001). Are Stock Prices Predictable? Retrieved from http://www.yorku.ca/ptryfos/randwalk.pdf