Financial Statement Analysis using Unsupervised and Supervised Learning

George Trieu[1], Jackson Kehoe[2], Alexia Tecsa[3], Raisa Sayed[4], Nicolas Wills[5]

QMIND – Queen's AI Hub
Queen's University, Kingston, Ontario K7L 3N6, Canada. Queen's

1 e-mail: g.trieu@queensu.ca
2 e-mail: 17jpk3@queensu.ca
3 e-mail: alexia.tecsa@queensu.ca
4 e-mail: raisa.sayed@queensu.ca
5 e-mail: 17nvw@queensu.ca

Abstract: *With the growing number of self-investors in the financial markets today, there must be more tools to help investors make informed decisions. The goal of this project is to group similar companies together and perform basic peer analysis and comparison on these companies. This is achieved by using various clustering methods to group companies together, and finally, perform analysis on these clusters through supervised learning and Shapley values. The attributes of each company selected to perform clustering were important accounting ratios to determining a company's success - both financially and on the stock market. The final model uses affinity propagation clustering and produces thirteen final clusters.*

---

1. INTRODUCTION

1.1 Motivation

In 2020, more than 2.3 million Canadians opened investing accounts [1]. This trend was found all throughout North America, as the lead personal investing platform, Robinhood, alone saw 13 million new traders in the past year [2]. As a result of these trends, there has been a growing concern as to the responsibility of non-professional traders, and the lack of knowledge behind their investment decisions. This has resulted in many Canadians being placed in positions of high financial risk. As interest in stock investments continues to grow among the public, it is important to provide stock analysis tools to create more informed decisions, thus placing the public at less financial risk.

1.2 Related Works

There have been numerous attempts to build comparative financial analysis tools for equities in the stock market using machine learning techniques. In 2015, the Marbaselios College of Engineering developed a clustering and regression model for stock prediction [3]. The research developed demonstrated that partitioning-based clustering performed better than density-based clustering and hierarchical-based clustering. Another similar project was developed by the Intel Institute of Science [4]. In the project developed by Intel, hierarchical agglomerative and recursive k-means clustering was effectively used to predict the short-term stock price movements after the release of financial reports.

1.3 Problem Definition

The problem tackled in this project is to develop a method to compare and group different publicly traded companies. This was achieved by employing various clustering models to separate and classify companies; as well as by determining the key financial metrics in each grouping.

2. METHODOLOGY

This project was completed in three phases: data preparation, clustering technique experimentation, and supervised learning.

2.1 Data Collection and Preparation

A dataset from kaggle.com containing 200+ Indicators of US Stocks from 2014-2018 [5], was used for this project. This dataset was then prepared for use by eliminating blank "Not a Number" (NaN) value rows and narrowing down the number of attributes contained within the dataset. The accounting ratios (attributes) used were those that investors would most commonly use to assess the financial performance of any company as this would allow us to better analyze firms across industries. These would include earnings per share and the price-earnings ratio which most investors are typically concerned with. The data was also normalized prior to clustering.

2.2 Clustering Models

Once the dataset was cleansed, various clustering techniques were applied to it such as k-means, DBSCAN, spectral, agglomerative, Gaussian mixture, and affinity propagation. Each clustering technique was then compared to one another through the quality of the clusters generated. For example, DBSCAN yielded most of the companies forming in one singular cluster, which is not useful for the purposes of this project.

2.3 Supervised Learning

Once the most effective clustering technique was selected, the Random Forest supervised learning was trained using the cluster number as the target attribute, to learn more about the feature importances of the clusters. Shapley values were also used to further explain the significance behind each attribute in the model.

3. RESULTS AND DISCUSSION

3.1 Clustering Attributes

The attributes (accounts/ratios) of the dataset were truncated to only keep the ones that were useful. The final list of accounting ratios used for clustering is shown in Table 1.

Table 1: Accounting Ratios used for clustering the data.

Net Cash Flow/ Change in Cash	Average Payables	Average Receivables
Current Ratio	SG&A to Revenue	Days of Payables Outstanding
Days of Inventory Outstanding	EBIT per Revenue	Debt to Assets
Debt to Equity	Payout Ratio	Return on Equity
R&D to Revenue	PE Ratio	Dividend Yield

After cleaning the dataset and isolating for the above attributes, 3568 companies remained for use in the clustering process.

3.2 Clustering Methods

The clustering methods experimented with were k-means, DBSCAN, agglomerative, spectral, Gaussian mixture and affinity propagation. Of the following techniques only Gaussian mixture and affinity propagation yielded distinct clusters between the companies due to the large variation within the normalized data. The unsuccessful techniques yielded most of the companies forming in the same cluster or the marking of most data as noise.
The affinity propagation clustering method works by comparing the different data points within the data to each other using matrices. When two points attributes are similar enough, they form a criterion matrix for the newly formed cluster which other data points must satisfy to join this cluster. This method eliminates the need to specify the number of clusters and different numerical metrics.
The affinity propagation clustering formed 46 clusters and had a silhouette coefficient score of 0.367. Some of the clusters formed with limited data points which indicated noise within the data. These clusters were filtered out and resulted in 13 distinct clusters.

```
Out[16]: 2     1498
         1      931
         5      441
         6      206
         4      100
         8       97
         0       93
         7       52
         3       21
         10      16
         11      11
         9       11
         12      10
         Name: cluster, dtype: int64
```

Figure 1: Results from affinity propagation clustering. The cluster number is on the left and number of companies in each cluster is on the right.

3.3 Random Forest and Shapley Values Analysis

After the formation of the clusters, they were analyzed using a combination of random forest and Shapley values to identify the dominant features. The random forest gave a very high accuracy score of 94.74% in identifying which cluster each company belonged to. The most influential features in the formation of clusters were debt to assets, current ratio, and debt to equity.

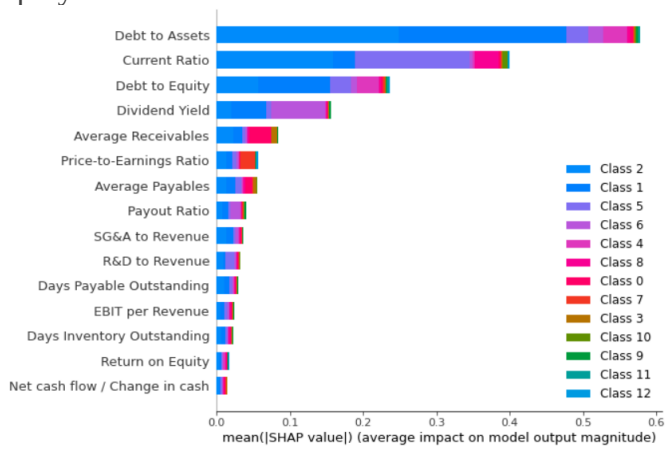

Figure 2: Contribution of features to the formation of different clusters.

Shapley values incorporate the model produced by the Random Forest algorithm to measure the contribution of a feature in each cluster individually. This is done using coalition game theory, by measuring the importance of each attribute to the predicted value [6]. An example can be seen in cluster 5, where most pharmaceutical and research companies were grouped.

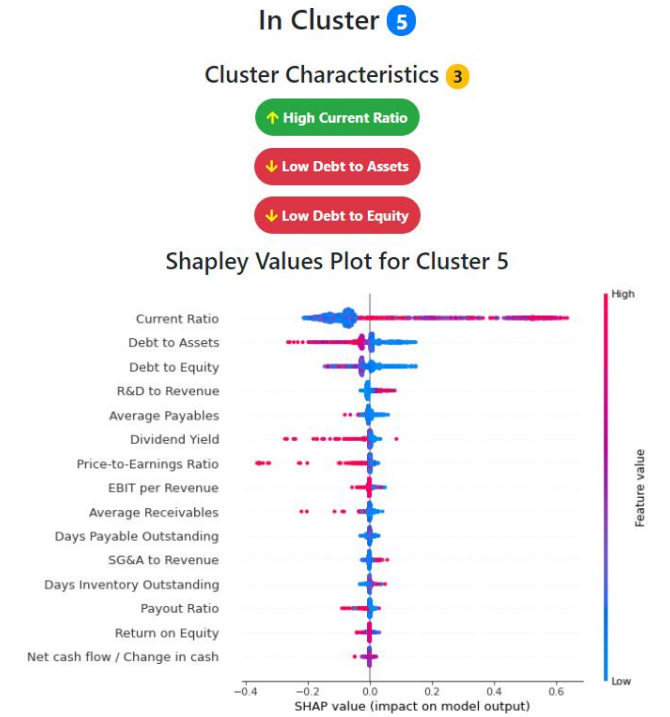

Figure 3: Sample results for cluster 5.

4. CONCLUSIONS AND FUTURE WORK

Without time constraints, further expansions to the project can be considered.

The first consideration would involve examining the companies within each cluster at a far more granular level to get a better understanding of their financial structures and the industries they operate in. This would allow the algorithm to further refine each industry cluster and identify closer similarities between companies within the same cluster.

The second application would be the analysis of the relative importance and impact of qualitative financial information on investment decisions. This analysis relies on non-quantifiable information such as management expertise, industry cycles, the strength of research and development, and labor relations [6]. Natural language processing can be applied to analyze textual data taken from management letters, financial statement notes, and other disclosed company announcements.

REFERENCES

[1] Global News, "Canadians opened 2.3 million DIY investing accounts in 2020. Should you?," 11 February 2021. [Online]. Available: https://globalnews.ca/news/7631776/diy-investing-canada-iiroc/.

[2] CNBC, "How Robinhood and Covid opened the floodgates for 13 million amateur stock traders," 7 October 2020. [Online]. Available: https://www.cnbc.com/2020/10/07/how-robinhood-and-covid-introduced-millions-to-the-stock-market.html.

[3] T. M. B.S. Bini, "Clustering and Regression Techniques for Stock Prediction," *Procedia Technology,* vol. 24, pp. 1248-1255, 2016.

[4] N. G. B. S. M.S. Babu, "Clustering Approach to Stock Market Prediction," *Int. J. Advanced Networking and Applications,* vol. 3, no. 4, pp. 1281-1291, 2012.

[5] N. Carbone, "200+ Financial Indicators of US stocks (2014-2018)," 2020. [Online]. Available: https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018.

[6] G. B. G. L. Huy Quan Vu, Data Mining Applications with R, 2014.

[7] T. Smith, "Qualitative Analysis," 7 March 2021. [Online]. Available: https://www.investopedia.com/terms/q/qualitativeanalysis.asp.