Project 1
MATH5430
Machine Learning for Finance
Apr 18th, 2022
Xingyu Lan

In this project, our goal is to determine whether financial statements help investors decide which stock to buy by giving us a higher return. Traditionally, investors may consider financial reports as a reasonable source of information to identify stock potential or company potential. But nowadays, given another dataset. The time used in making a decision is getting shorter, and these data are usually updated at a slow pace causing them to be overlooked. In this project, we are trying to use only financial data in the balance sheet to see how the stock performs. Other ways were left out because they are usually quite helpful and don't require a test like this. Therefore, machine learning is used to classify stock by utilizing stock's financial data and observing differences between groups to achieve this goal. Hopefully, the difference will be clear, and we can understand the difference between groups; then, one group may have the property investor desire or be different from the other stock. Separating the stock into groups wasn't the primary goal of this research. The goal is to determine how stocks are different, simply splitting them up wasn't helpful in our test.

First of all, we use a dataset from Kaggle; the dataset contains the following, the financial data from 2012 to 2016 of S&P500 company, the stock price data, and a small description of that data. Then, we try to use the k-mean classifier to take stocks into groups. After we deleted some rows containing missing data points and rename them to pretreat the data. Also, we have generated some custom data from the original data, such as the change in price per year, to suit our annual base dataset better.

We use the following parameters in the k mean model ;["TickerSymbol","year","accounts payable","accounts receivable","gross profit","Liabilities","NetCashFlow","operating income","total assets","TotalEquity","TotalLiabilities","TotalLiabilities&Equity","TotalRevenue","EarningsPerShare"].The last one were left out to to some of the column are left out. The dataset has also been standardized to suit our results better. And Ticket symbol and year were kept as indexes or labels for further use, So the remaining data are put in the k-mean cluster.

Both Elbow Curve Method and Silhouette analysis were used to determine the ideal cluster. They both agreed that two should be the number of clusters due to elbow point occurring on two and the highest silhouette coefficient values. Then, the actual k mean was run. But before we employ such a method on our test dataset, the result of the training dataset should also be considered; there is a bit of randomness, but one group is always significantly smaller, around 30 of them. Those companies are well-established and usually well-known firms. But such data wasn't valuable enough because any investor can look this up and tell which stock is from the most well-established company.

And by some simple statistical analysis, the smaller group may always have a lower return price on stock for next year and be indifferent in the current year; the result may not be evident due to such a small dataset. And their relationship continues as we can apply the same model to the training dataset. Still, again, it remains unclear how strong the relationship holds due to such a small dataset. Also, if we look at the graph, it is challenging to tell if there is any difference between the two groups. The result suggests looking at the balance sheet may not be worth the time if the goal was having a higher return on the stock price.