Jibo Wen
July 26th, 2017

Definition

Project Overview

Stocks are typically categorized by sectors, such as energy, financials, technology, healthcare etc, or by their market capitalization, such as S&P 500, S&P MidCap 400 or S&P SmallCap 600. Different segmentation methods will present different aspects of stock properties and behaviors. However, from time to time, stocks of different capitalization or sectors can have different growth rates (or decline rates) and volatility, creating a problem for investors who are looking for similar stocks based on their sectors or sizes.

The purpose of this project is to apply unsupervised learning to the stocks trading on NYSE and NASDAQ, and find an underlying clustering mechanism based on their financial conditions and growth rate. Such clustering mechanisms can present the stock market to investors from a different perspective, in which stocks exhibit similar financial conditions in each cluster, therefore, are likely to show similar performance. This new clustering mechanisms will also be compared to an existing mechanism such as grouping by sectors to test which methods yield better results based on the homogeneity of the stocks in each group.

Problem Statement

The goal of this project is to find more fundamental ways to group stocks, so that stocks in each group will exhibit similar financial standings and behavior. In order to find a stock segmentation scheme, there are two problems need to be addressed:

1. Identify relevant features to use for clustering
2. Find a clustering method based on the characteristics of the features

The strategy for solving these two problems will involve the following:

1. Retrieve NYSE and NASDAQ stock data from Yahoo Finance
2. Preliminary data exploration to understand the characteristics of the dataset
3. Data preprocessing including removing incomplete data and outliers and feature scaling
4. Apply Principal Component Analysis (PCA) to identify the most relevant feature combinations and visualize the dataset.
5. Apply unsupervised clustering algorithms including K-Means and Gaussian Mixture Model
6. Compare the performance of both models to existing segmentation methods based on their silhouette score

The final application of the project is to relabel all the stocks in the study based on their clusters, and offer a meaningful interpretation for each cluster.

Metrics

Silhouette score is a common way to measure how closely are the data in each cluster are grouped. It is the average of the silhouette value of all the data point. The silhouette value a data point is calculated by:

$$s_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

Where:

a_i: average dissimilarity of point i with respect to other data points in the same cluster
b_i: minimum average dissimilarity of point i with respect to data points in another cluster

In this project, Euclidean distance will be used to measure dissimilarity because stocks with larger Euclidean distance are likely to be more different than those with smaller distance (or similar values of their features). The silhouette score will range between -1 to 1 by definition. The closer it is to 1, the better the clustering scheme, as the data points are more similar to the data of their own clusters than others. Conversely, the closer it is to -1, the poorer the clustering scheme.

Analysis

Data Exploration

In this project, stock financial statistics are retrieved from Yahoo Finance[1]. Although many features are available, only 8 of which are selected for this study. They are:

1. Price Earnings Ratio (PE)
2. Price Sales (PS)
3. Price Book (PB)
4. Short Ratio (SR)
5. Dividend Yield (DY)
6. Price Earning Growth Ratio (PEGR)
7. Percent Change from Year High (PCYH)
8. Percent Change from Year Low (PCYL)

Feature 1-6 will be based on the most recent quarter as of June 2nd, 2017; feature 7-8 will be based on the year to date data as of June 2nd, 2017. Out of over 4000 stocks listed on NYSE and NASDAQ, only 2869 stocks whose financial information is complete are chosen for this project. A sample dataset is presented in Table 1.

Symbol	DY	PB	PCYH	PCYL	PE	PEG	PS	SR	SECTOR
NVDA	0.41	13.42	-0.0237	2.1824	47.61	3.68	10.91	1.41	Technology
JPM	2.34	1.32	-0.0917	0.4962	13.14	1.58	3.29	2.13	Finance
ADES	0.00	2.46	-0.2278	0.5233	2.09	0.07	5.88	5.57	Basic Industries

Table 1: A sample of the dataset with Nvidia, JP Morgan, and Advanced Emissions Solutions

The reason these features are chosen is that they are all ratios rather than absolute values, as absolute values may not represent a stock accurately. For example, a stock with a lower price is not necessarily cheaper than a stock with a higher price. The value of a stock is more accurately represented by their PE or PB ratios. Although there are many features that are ratios, only the 8 are chosen because they are widely available and commonly used to evaluate stocks. Of these 8 features, Feature PE, PS, PB, DY and PEGR offer a financial perspective of a stock, Feature SR a trading perspective, and Feature PCYH and PCYL a growth perspective.

Exploratory Visualization

A summary of the dataset is shown in Table 2.

[1] Details of data retrieval can be found in iPython notebook Stock_Data_Retrieval.ipynb

	DY	PB	PCYH	PCYL	PE	PEG	PS	SR
count	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000
mean	1.923789	5.262124	-0.137261	1.036934	50.919360	4.647964	3.663936	5.036375
std	2.590263	32.518166	0.131984	33.707553	219.507561	140.381934	14.516095	5.342447
min	0.000000	0.070700	-0.942500	0.000000	0.075000	-4098.680000	0.010000	0.000000
25%	0.000000	1.360000	-0.188500	0.173100	15.500000	0.000000	1.010000	1.830000
50%	1.250000	2.130000	-0.103400	0.315600	21.670000	1.300000	2.250000	3.470000
75%	2.720000	3.760000	-0.041500	0.518400	34.030000	2.290000	4.192800	6.510000
max	35.740000	1191.530000	0.295000	1805.749800	6385.000000	4299.500000	529.420000	58.350000

Table 2: Statistical summary of feature space

According to Table 2, there are a few observations can be made:

1. Most features have extreme values that are far above or below their 25, 50 and 75 percentiles, which can cause inaccurate clustering.
2. PCYH is negative whereas all other features are positive. PCYH should be negative by definition. Any positive PCYH should be removed.
3. Some features (such as PE, PEG) have a larger range and variability than other features (such as PCYH, PCYL) by the way they are defined. Rescale these features to similar range will help PCA identify more important dimensions.

In addition to a statistical summary, a histogram of each feature is shown in Figure 1. The distribution of each feature can be visualized and it is clear that outliers have heavily skewed the distributions. They need to be removed to allow a more truthful visualization of the feature distributions.

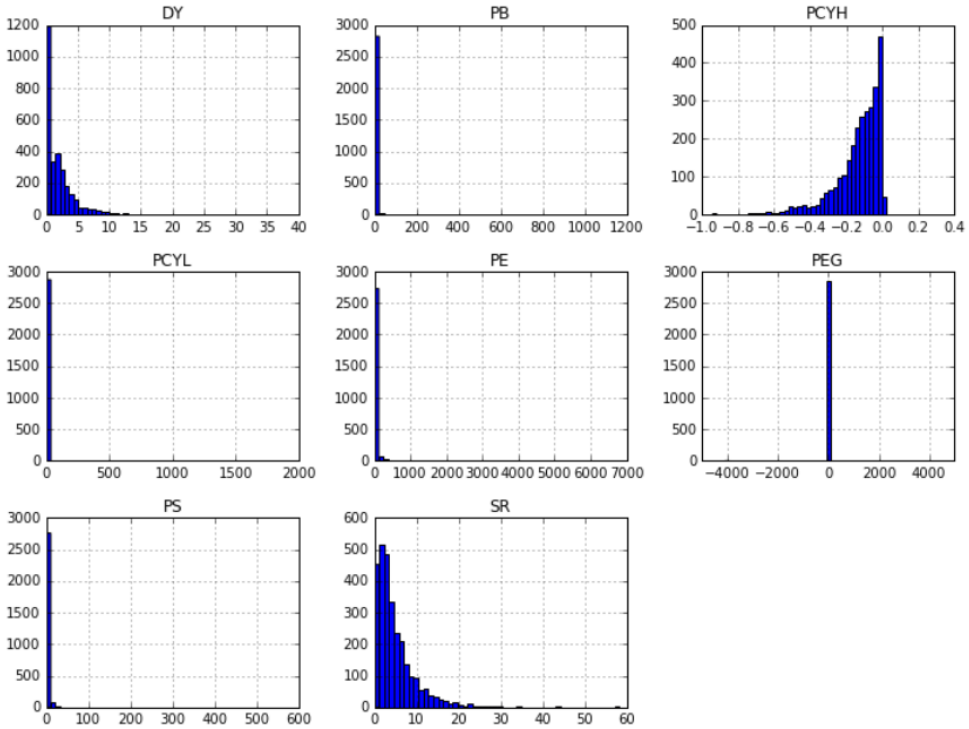

Figure 1: Histogram of all features

Algorithms and Techniques

As features might correlate with each other, some features might be more relevant than others. Feature relevance can be determined based on how well can they be predicted from other features. The feature that can be well predicted is less relevant as its information has already been contained in other features. A linear regressor, a decision tree regressor and a SVM regressor will be used for prediction and Coefficient of Determination (R score) will be used for measuring feature relevance. The higher the R score, the less relevant the feature will be.

Once feature relevance is determined, PCA can be used to compress the dataset to lower dimensions, allowing to visualize the dataset as well as identify the most relevant features or feature combinations that can explain the most variation of the dataset. However, such features or feature combinations might not exist. In other words, PCA might fail to reduce the dimensionality of the dataset while reasonably keeping their variation.

If the PCA succeeds, K-Means and Gaussian Mixture clustering algorithms will be applied to the reduced dataset. Otherwise, they will be applied to the original dataset. These two unsupervised learning algorithms will cluster the data points into several groups. The number of groups can be determined by the one with highest silhouette scores. The clustering models will then be compared to the benchmark model and might offer fresh and meaningful perspectives to understand the stock market.

Benchmark

The benchmark model for this project is to group stocks by their sectors, which is a common and natural way to group stock. There are 11 sectors in the dataset such as Finance, Consumer Service, Technology etc. We can calculate the silhouette score of this group scheme and compared it to the proposed clustering models. The benchmark result will be calculated in later sections, after the data has been preprocessed.

Methodology

Data Preprocessing

Based on the observations from Data Exploration and Visualization, there are a few steps needed to preprocess it:

1. Drop any duplicate data points.
2. Remove outliers based on the inter-quartile range (IQR). Typically for normally distributed data, 1.5 times the IQR above 75 percentiles or below 25 percentiles will be used as cut offs for outliers. However, due to the skewness of this dataset, a more tolerant range will be used.
3. Remove data points with positive PCYH, and take the absolute value of PCYH to convert them all to positive for consistency with PCYL.
4. Scale the features to range 0 to 1.

The preprocessed data statistics is summarized in Table 3, and a matrix scatter plot of any two features is shown in Figure 2, with a histogram of each feature on the diagonal. Since the data has excluded outliers and scaled, their distribution can be clearly visualized, and they are all unimodal distribution skewed to the right. The scatter plots don't show strong linear relationship between any two features. Nevertheless, some features are slightly inversely related such as DY and SR. This is consistent with intuition as stocks with higher dividend yield are less likely to be shorted because they are making more profit, and vice versa.

	DY	PB	PCYH	PCYL	PE	PEG	PS	SR
count	2674.000000	2674.000000	2674.000000	2674.000000	2674.000000	2674.000000	2674.000000	2674.000000
mean	0.100765	0.115432	0.146516	0.106703	0.135142	0.486524	0.085996	0.095996
std	0.130037	0.120168	0.138981	0.094089	0.129229	0.061703	0.087501	0.099092
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.047552	0.045522	0.046486	0.069218	0.450791	0.028106	0.035272
50%	0.068002	0.075228	0.111312	0.084872	0.096552	0.479679	0.061973	0.066349
75%	0.142857	0.135133	0.200419	0.138455	0.145927	0.500165	0.114882	0.124309
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Table 3: Statistcal summary of the preprocessed dataset

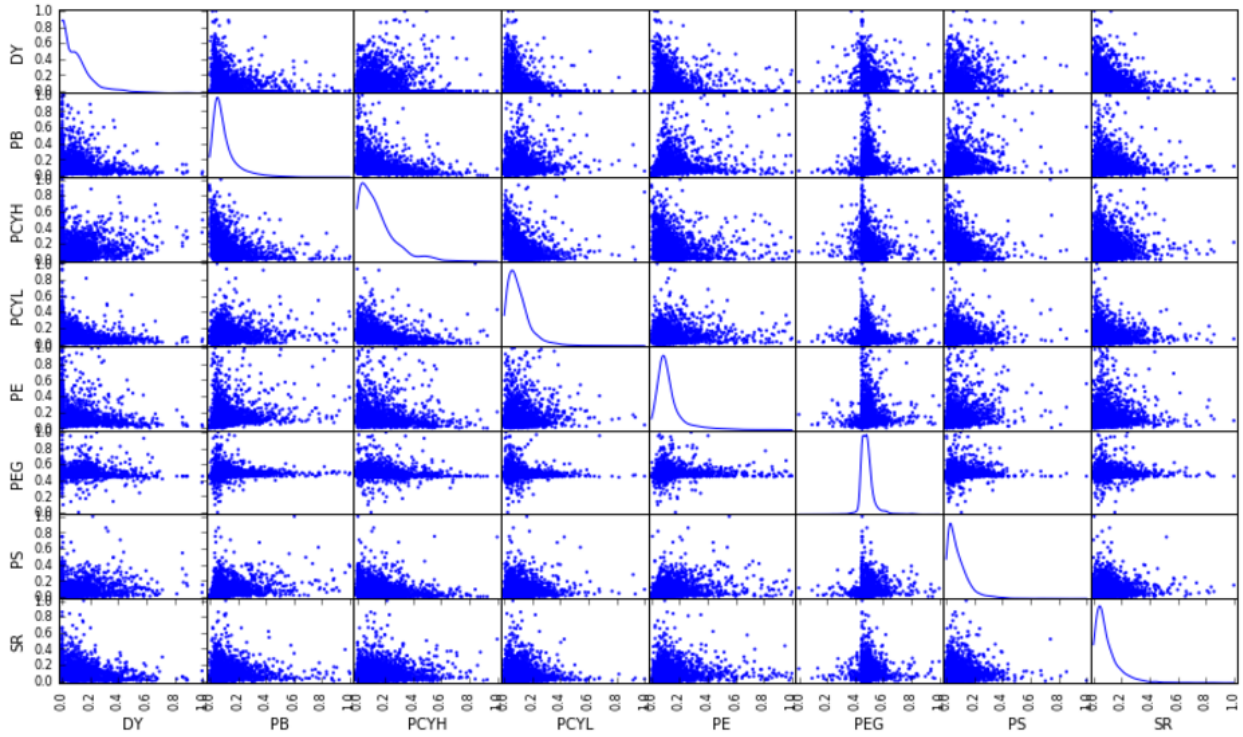

Figure 2: Matrix scatter plots of any two features, with histogram of each feature on the diagonal

Implementation

The implementation process can be broken down into three parts: Feature Relevance, PCA, and Clustering Analysis.

Part 1: Feature Relevance

Feature relevance study include the following steps and the results is summarized in Table 4:

- Drop a feature from the dataset for later use of other features to predict its value
- Split the dataset into training and test set
- Train a regressor (Linear Regressor, Decision Tree, SVR) on the training set
- Predict the value of the test set
- Calculate the coefficient of determination

	Linear Regression	Decision Tree	SVR
DY	0.0717958	-0.466577	0.0597555
PB	0.158387	-0.654897	0.103447
PCYH	0.100279	-0.228194	0.148848
PCYL	0.0581436	-0.706022	0.0150209
PE	0.147818	-0.312307	0.124602
PEG	0.0398144	-1.16323	-0.166757
PS	0.222831	-0.752889	0.081005
SR	-0.0319298	-1.47558	-0.097128

Table 4: Coefficient of Determination for feature relevance study

According to Table 4, the linear regressor slightly outperforms decision tree and SVR (with a radial basis function kernel) regressors. However, the highest coefficient of determination is 0.223 when predicting Price to Sales using linear regression, meaning only 22.3% of the variation in PS can be predicted by other features. This result is consistent with Figure 2, in which no two features exhibit strong linear relationship.

Part 2: Principal Component Analysis

The Principal Component Analysis include 2 steps:

- Apply PCA to the data set and identify the composition of principal axes.
- Reduce the dataset into 2 dimensions for visualization.

The composition of the principal axes of the transformed dataset is shown in Figure 3. A few observations can be made. First, only 44.44% of the variation are captured by the first two principal dimensions, and 61.16% by the first three principal dimensions. Although some dimensions are more important than others, there aren't any dominant dimensions that can capture the majority of the variation. This is consistent with the result of Feature Relevance study, as no single feature can be well predicted by others. Second, performing clustering algorithms on reduced dataset (of two or three dimensions) might not be relevant, as almost half of the variation is absent. However, we can still use the reduced dataset with two dimensions to visualize the clustering process. The final clustering model should be performed on the original dataset. Last but not least, the three most distinctive features are PCYH, PE and DY, as they are the tallest among the first three principal axes. Typically, stocks with higher PE will have lower DY, as they are still growing rapidly. Stocks with lower PE are mature companies that usually pay more dividend to attract investors.

The projected data points on the principal plane along with axes of original features are shown in Figure 4. The reduced two-dimensional dataset exhibits triangular shape, with PCYH, PE and DY axes pointing at each corner, indicating three important aspects of stocks, price change, earning and profit. A natural starting number of clusters would be three, and it can be further adjusted in Refinement section.

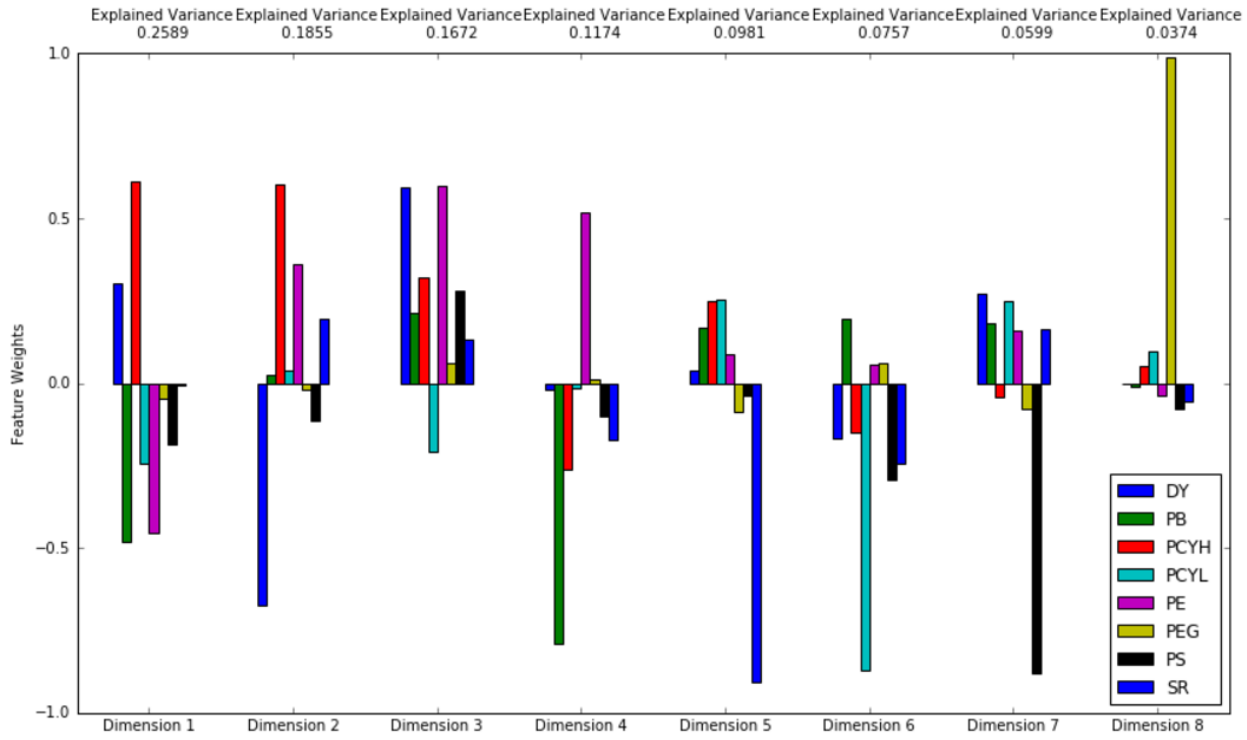

Figure 3: Principal component composition of the transformed dataset

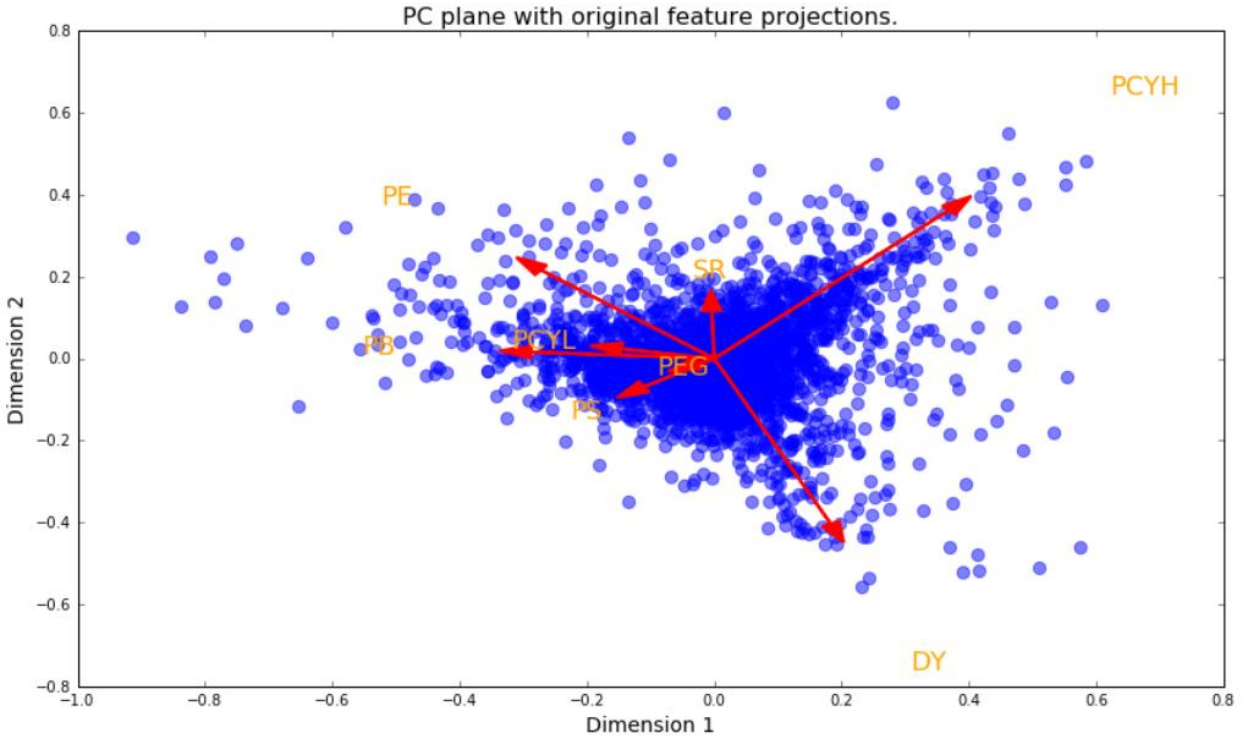

Figure 4: Principal component plane with projections of original features

Part 3: Clustering

Gaussian Mixture and K-Means clustering mechanisms will be used on the reduced dataset with two dimensions and the original dataset with eight dimensions, with three clusters to start with. The Silhouette scores of Gaussian Mixture and K-Means are summarized in Table 5. The results of the reduced dataset are shown in Figure 5 and Figure 6.

In Figure 5, the Gaussian Mixture model breaks the data into three groups, the green group tightly crowded at center, the yellow group spreading horizontally on the left (more variation along dimension 1), and the blue group spreading vertically on the right (more variation along dimension 2). The green group might be interpreted as an average stock group whose feature values are close to the average of all, the yellow group as stocks with high variation of PE and PB, the blue group as stocks with high variation of PCYH and DY.

In Figure 6, the three groups are formed by trisecting the triangular data shape into three corners, each of which represents a dimension of the original dataset according to Figure 4, which are PCYH, PE and PB, and DY.

Both model can be interpreted by grouping data along a major dimension of the dataset, but according to Table 5, the Gaussian Mixture model has a larger silhouette score than the K-Means has.

Silhouette Score	Gaussian Mixture	K-Means
Reduced Dataset	0.42	0.36
Original Dataset	0.26	0.25

Table 5: Silhouette scores of clustering results using 3 clusters

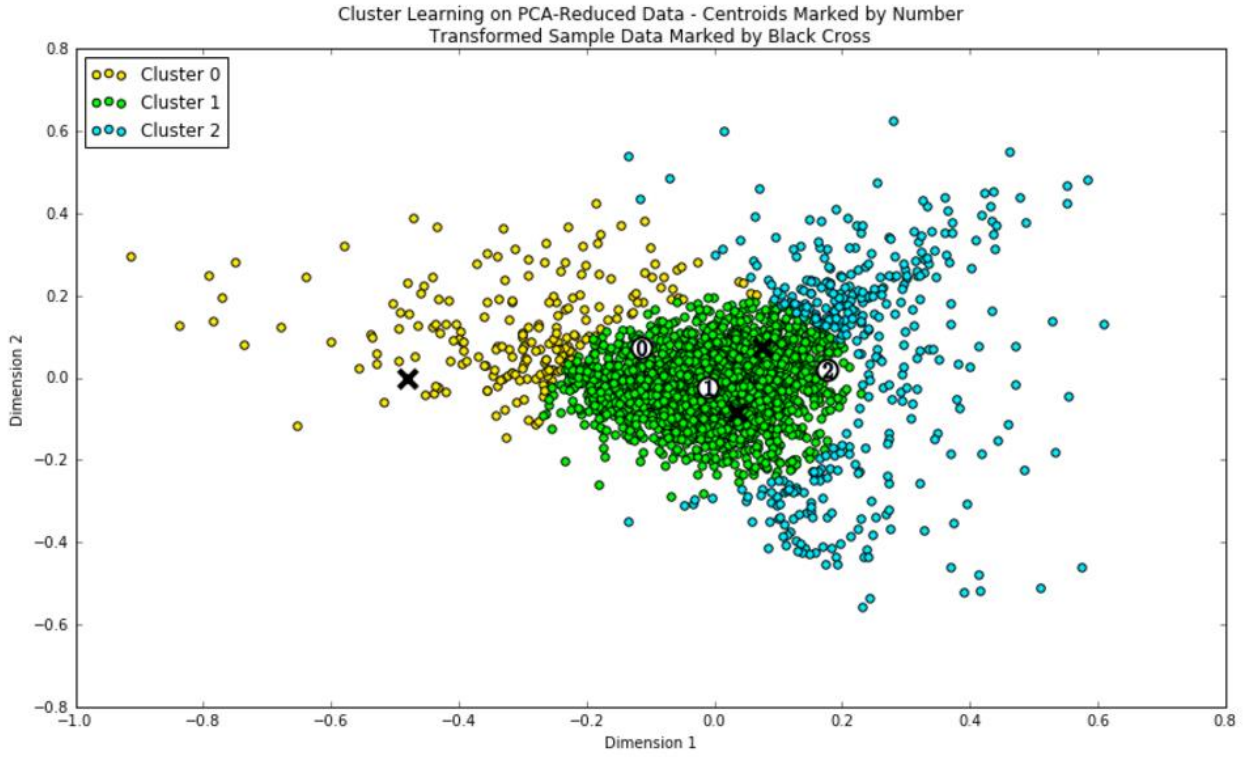

Figure 5: Gaussian mixture clustering on the reduced dataset

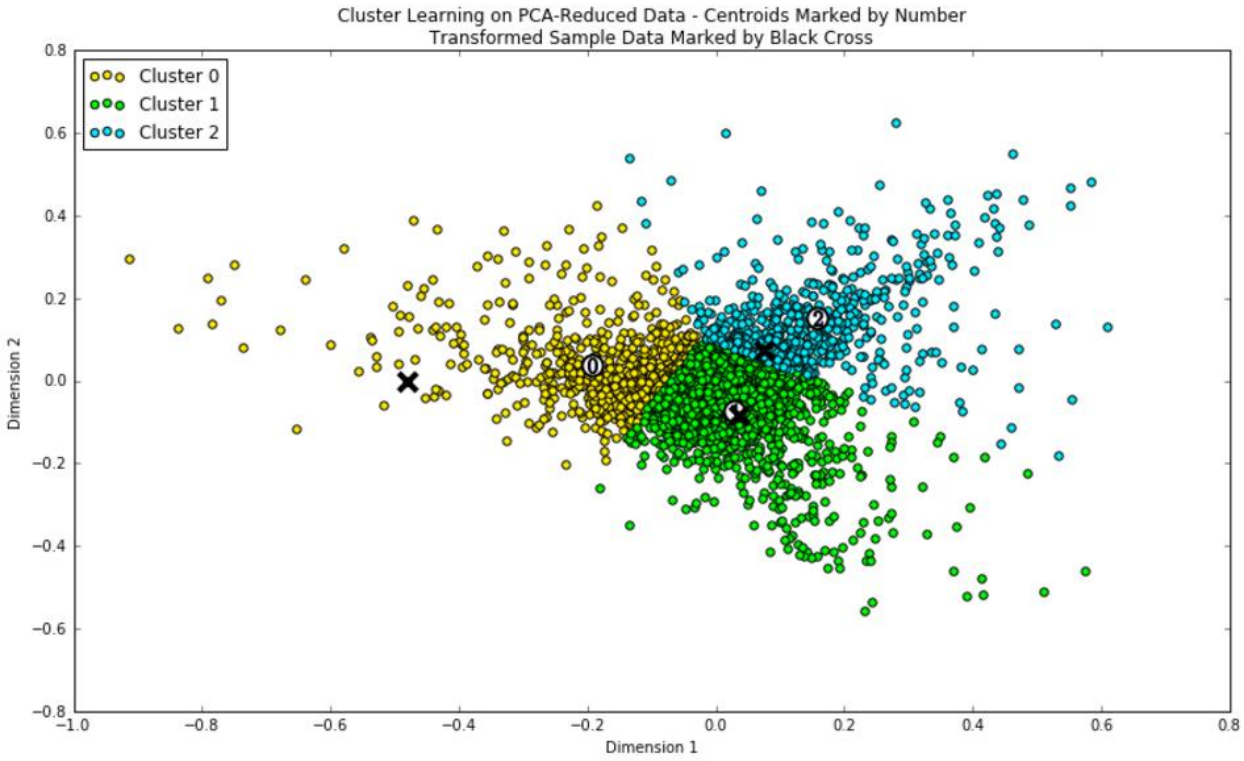

Figure 6: K-Means clustering on the reduced dataset

Refinement

Both the Gaussian Mixture and K-Means models can be refined by finding the optimal number of groups, which can be determined by their silhouette score. Silhouette scores of 2-12 clusters are calculated based on the clustering results of the reduced dataset and the original dataset, and are plotted in Figure 7 and Figure 8.

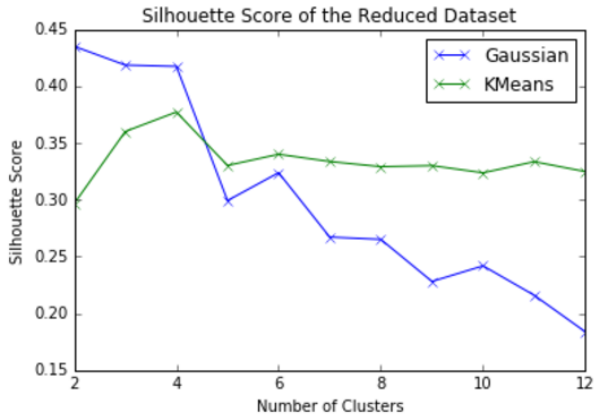

Figure 7: Silhouette score of clustering the reduced dataset

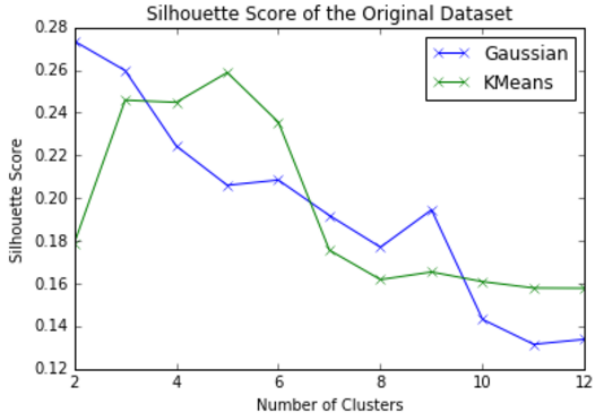

Figure 8: Silhouette score of clustering the original dataset

Since the final model will be selected based on the performance of the original dataset, the best silhouette score from Figure 8 is the Gaussian Mixture model with two clusters. Hence, it will be the model chosen to compete against the benchmark model.

Results

Model Evaluation and Validation

The final model is determined as Gaussian Mixture model with two clusters trained on the original dataset with eight dimensions. The reason it has to be trained on the original dataset is that PCA is not able to reduce the dataset into less dimensions while keeping reasonable amount of variation, therefore, the reduced the dataset with two dimensions is used only as a visualization aid. The reason for two clusters is that it is the one with highest silhouette score among all other models (K-Means or Gaussian Mixture) with different number of clusters.

Two tests have been conducted to evaluate the robustness of the model. The first one is to test the model on a random subset of 90% of the original data; the second is to test on the original dataset with a random noise in the range of 0 to 0.05, about 5% of the original dataset in terms of variation. Test 1 result is shown in Figure 9 and Test 2 Figure 10.

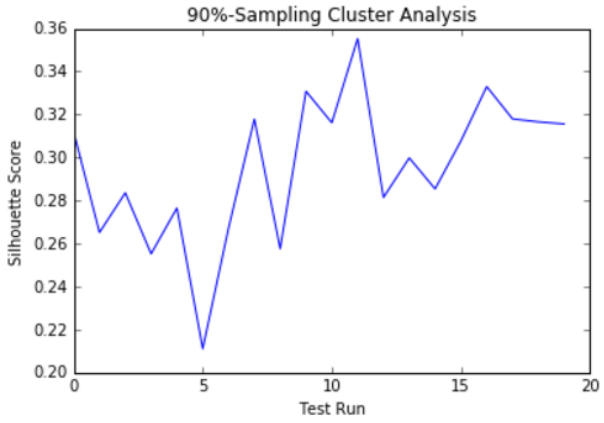

Figure 9: Sub-sampling test

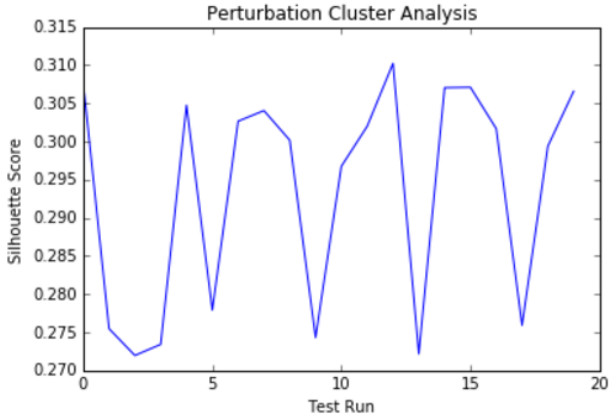

Figure 10: Random noise test

The silhouette score of the final model varies about 20% in both test, and the worst performance is about 0.21 in Figure 9. Based on Figure 9 and 10, the model is fairly robust for carrying out the clustering task as its performance varies only on a small scale when subjected to random noise or missing data. Because of the robust performance, the result can be trusted as the model is capable of clustering the original dataset into two groups with a silhouette score of 0.27. Although the silhouette score is far from 1, the model is reasonable because the dataset, according to Figure 2 and Figure 4, doesn't exhibit clear clusters. Therefore, creating more groups will only break the data in artificial ways that lead to worst performance as proved by Figure 7 and Figure 8.

Justification

The benchmark model is to cluster stocks based on their sector labels. The silhouette score of the benchmark model is calculated as -0.0746. The final model eclipses the benchmark model in clustering performance based on the silhouette score. The final model offers a better clustering scheme in the eight-dimensional dataset chosen for this project by grouping data that are more similar in their underlying

distribution. The main reason the benchmark model performs poorly is it divides the data into eleven groups, but according to Figure 4, there are not clear boundaries to separate the dataset into eleven groups. Hence, it fails to capture the underlying distribution of the features.

Conclusion

Free-Form Visualization

To understand why the benchmark model performs a lot worse than the final model, we can visualize the results of the benchmark model on the reduced dataset (from PCA). The result is shown in Figure 11.

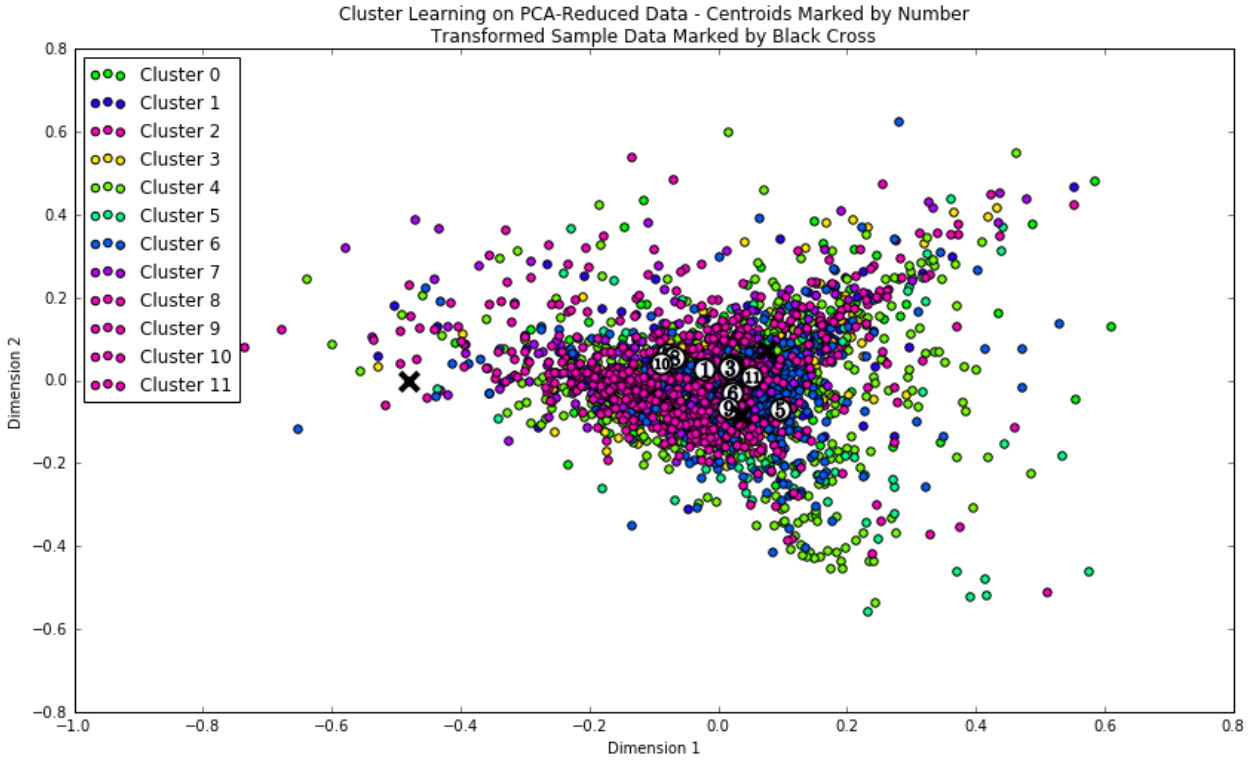

Figure 11: Cluster visualization of the benchmark model

As shown in Figure 11, there are eleven clusters. The data points of each cluster scatter across both dimensions without obvious pattern to justify. This is consistent with the intuition that stocks in the same sector can exhibit drastically different financial properties and performance.

Comparing to the benchmark model, the result of the final model on the reduced dataset is shown in Figure 12. It separates the data into two groups. The first group in green tightly sit in the center of the principal component plane, whereas the second group in yellow distributed across the plane. The first group might be interpreted as average stocks, a stock with average financial characteristics and performance. The second group might be interpreted as stocks with one or more outstanding characteristics in terms of financial or performance aspects. This is a very coarse clustering, but can be further refined with the addition of more features.

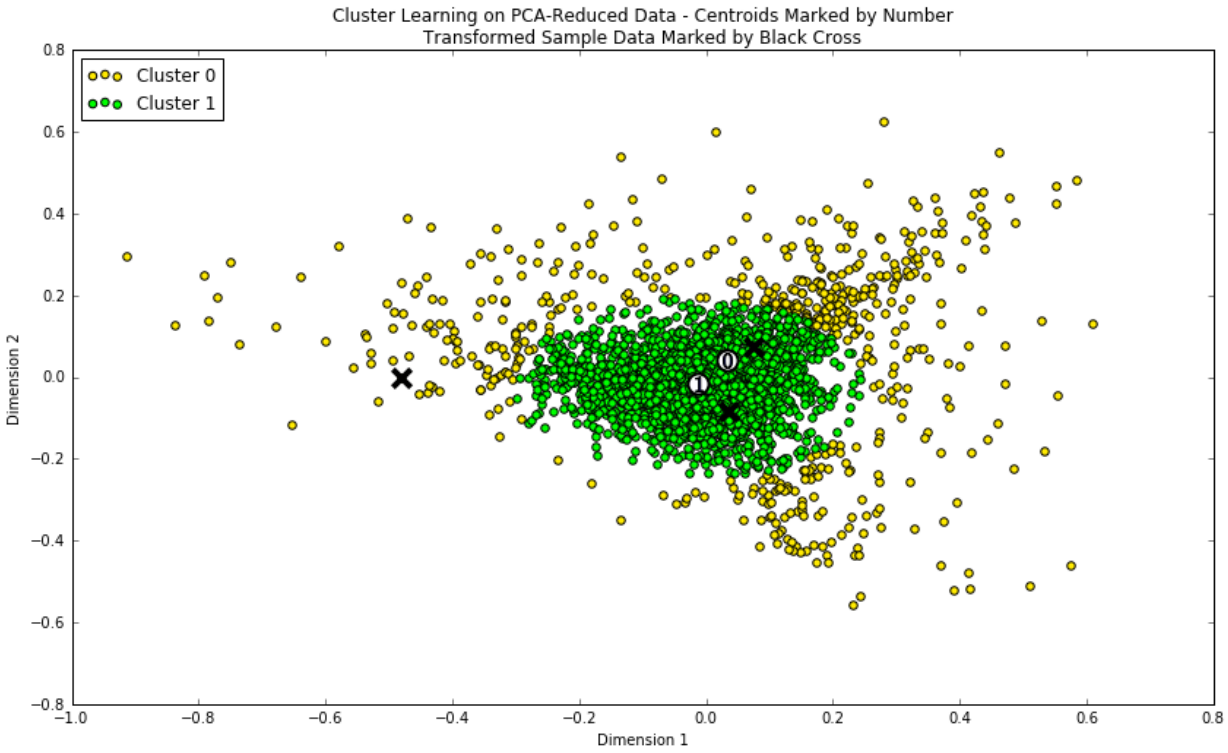

Figure 12: Clustering visualization of the final model

Reflection

A summary of the work flow of this project is shown below:

1. Defined a problem and a benchmark solution.
2. Scraped a relevant dataset from web.
3. The dataset was examined for potential preprocessing needs.
4. Preprocessed the data including outlier removal and feature normalization.
5. Quantified feature relevance by studying the correlations between them.
6. Transformed features using PCA and examined the transformed features and variation they explained. The PCA was not able to reduce the dataset into two or three dimensions while keeping reasonable amount of variation, therefore, it's used as tool to visualize the dataset by reducing it to two dimensions.
7. Clustered the reduced and the original dataset by using K-Means and Gaussian Mixture. Refined the models by determining the number of clusters that yield the best silhouette score. Select a final model that best clustered the dataset.
8. Compared the final model to the benchmark model.

The interesting part of this project is applying feature scaling prior to PCA analysis. I started the project without scaling any features, and found a lot of the variation was lost after dimensionality reduction by PCA. The most significant dimension would always be the feature with the largest range of values, such as PE, eclipsing other features. I tried to scale the dataset in many different ways, and get drastically different results. Feature scaling became the most crucial and difficult part of the entire project as it will influence every following step. The eventual method for scaling the feature is to normalize all the data in the range from 0 to 1. The reason is that many features are percentage value (such as DY, SR, PCYL, PCYH), and have already been in that range. Scale all the features into that range will eliminate the impact of absolute value, and emphasize the impact of variability.

The final model and solution does fit my expectation to a limited extent, as it is able to group stocks that are similar to each other in terms of the 8 features selected much better than the benchmark model is. However, due to the dataset itself may not be so separable, or the absence of observable clusters existing within the dataset, the model may have some limitations.

Improvement

To find potential improvement, we need to understand the limitations of this study. The foremost important factor is dataset. In this project, only eight features are selected to represent a stock. This representation is by no means comprehensive. In reality, a stock could have hundreds of features from technical, financial and other perspectives. How to select a representative feature space is beyond the scope of this project. However, if a more relevant and informative feature space can be determined, the clustering algorithms may yield more interesting results.

On the algorithm side, only two clustering algorithms have been explored, K-Means and Gaussian Mixture. They are centroid based and distribution based clustering algorithms respectively. There are also connectivity based and density based algorithms. Depends on the dataset and distribution, any of the clustering algorithms could be more appropriate than the others. Therefore, carefully select an algorithm based on the properties of the dataset could potentially improve the performance.