ECE 498 Project Proposal Report
4-credit hours, Spring 2020

Title: Feature generation for portfolio diversification

Team: gowtham4, mananm2, somani4

Problem Statement:

The primary task is to cluster the companies based on key features (Feature Engineering). Identifying these features to efficiently perform clustering is a major challenge. Evaluation criterion for the efficacy of the generated features is comparison of predicted sectors with Global Industry Classification Standard (GSIC). This will help verify the accuracy of the features generated and tuning these features iteratively to get better accuracy for further analysis. The next problem is to maximize the returns for an individual by analyzing the stocks of the companies and performing stock portfolio diversification using the obtained domain knowledge

Motivation:

Personal finance is an important aspect in every individual's career. 'Beating the market', as is often discussed, is not as imperative for personal finance as 'following the market' to make smart investment decisions. A lot of work has gone into predicting future stock performance based on its past indicators. However, our goal is to explore how seemingly un-correlated industries affect each other's stock prices. We intend to identify the key features that form these correlations. We intend to take an S&P 500 dataset and cluster like stocks together based on these features using clustering algorithms learned in class and perform portfolio diversification. In part we are trying to level the market indexes.

Description of the Data:

1. **https://www.kaggle.com/camnugent/sandp500**
2. **https://en.wikipedia.org/wiki/List_of_S%26P_500_companies**

We are using (1) as our primary dataset. This dataset is a time series which consists of 5 numerical columns and 1 string data for each row. Open, High, Low, Close, Volume traded as the numerical columns and stock ticker are available for each company on the S&P 500 index for the past few years. We will generate our features from this data. This task is elaborated in the solution section below.

Example: AAPL - Apple has this timeseries. And there are 500 such tables.

Date	Open	High	Low	Close	Volume
Start: 2013-02-08	67.7	68.4	66.9	67.8	158,168,416
End: 2018-02-07	168.1	163.4	159.1	159.5	51,608,580

Data set (2) will be used to extract GSIC and group by sector and/or sub industry. This will act as the supervisory “y” values of the cluster.

Related Work & Plan of Action:

Earlier articles explored feature generation and clustering. Our novel additions in this project are generating unique features and using thus obtained domain knowledge in the field of personal finance. We haven’t seen much research in this direction. Alpha is used in finance as a measure of performance, indicating when a strategy or portfolio has managed to beat the market return over some period. We can use this to evaluate the performance of our portfolio.

We know there are 11 definitive sectors of the economy into which the S&P 500 stocks can be divided as per their weightings by market capitalization. The sectors and their distributions are as follows: Communication Services (9.9%), Consumer Discretionary (10.2%), Consumer Staples (6.7%), Energy (6.0%), Financials (13.7%), Healthcare (14.9%), Industrials (9.7%), Materials (2.5%), Real Estate (2.7%), Technology (20.8%), and Utilities (2.8%).

Given the data with only basic features like Open Price, Close Price, Volume Traded and High-Low Prices, we want to start with **feature generation** to find a set of ‘worthy’ features which can cluster stocks in the groups shown above. Finding these features will require domain knowledge. Some common indicators used in stock analysis, as mentioned below, can be tried initially:

- Moving Averages (linear and exponential)
- Supports
- Resistances
- Normalized Daily, Monthly and Annual Ranges
- Bollinger Bands

The purpose of this analysis is to find underlying features (which may not be apparent from the initial data) that link similar stocks together. Further, different **clustering algorithms** (K-Means, GMM, Neural Nets) can be applied to different sets of features to find the best algorithm-feature match. Once we have a feature set, **Principal Component Analysis** (PCA) can be performed to reduce feature size based on feature variabilities. PCA is particularly important as 2 or more features (like linear and exponential moving averages, for instance) might be significantly correlated. All the above analysis can be performed in multiple time scales *viz.* Daily, Weekly, Monthly, Quarterly, and Yearly. We presume that most insight is available at the Daily and Quarterly time scales.

Finally, the learned parameters (features) from the previous clustering analysis, we intend to generate a new function which tells us whether adding a stock to a current portfolio is advisable or not (and to what extent).