ECE 498 Project Progress Report 1
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.

Progress Made:

Feature Engineering:

1. **The Need For Features:**

   We are drawing upon prior literature on Technical Analysis on stocks as opposed to fundamental analysis where we don't necessarily look at daily stock transaction data. Each stock market data entry consists of the following base features: "open, high, low, close, volume" for a given day. These are the fundamental features that any stock would have, but these alone are not sufficient to accurately study the trends and predict stock changes. Hence the need to explore, study and cleverly choose the more refined features. These refined features serve as a better criterion to study trends for a particular company. Comparing these features across all companies would lead us to identifying the similarities between the inter-company stocks, and hence, effective clustering into the industry sectors.

2. **Judicious Feature Selection:**

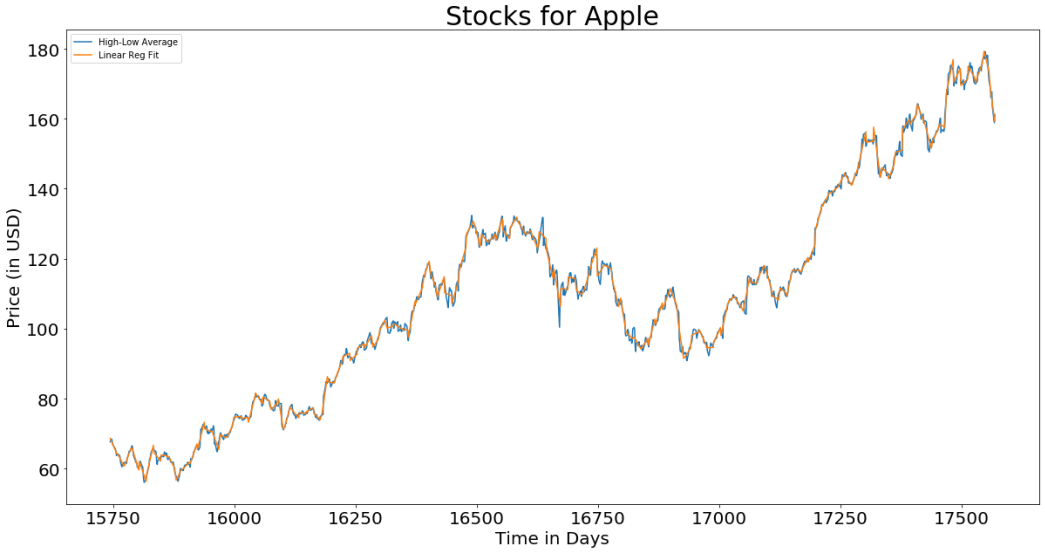
   These features (also called indicators) usually signify the volatility, momentum, spread and other details across multiple timeframes. A combination of these may yield useful clustering information. Upon investigation and gaining some domain knowledge, it was found that there are 110+ popular features which could be potentially used to study and predict the stock market for a particular company. The features (depending on their mutual similarities) were equally divided among the team-members (~35 features per person) and studied in-depth to understand what knowledge is gained by implementing these features. Depending on the amount and type of information conveyed by the features and comparative feasibility of implementation, roughly 10 features (per person) were narrowed down as the ones that can potentially connect all companies in a similar sector together.

   The features were selected based on what they 'indicate' for a stock. Different features have different utility, for instance Average True Range (ATR) measures volatility, which is useful when clustering stocks for a diversified portfolio. However, some indicators like Choppiness Index (measures direction of the intraday market) or Keltner Channel (used as a visual aid more than a stand-alone feature) had to be discarded, as no useful information was conveyed by them

for long term analysis. Most of the technical analysis indicators were developed for intraday and positional trading. But some of them can be repurposed to work for longer time frames as required in investing scenarios.

3. **Feature Implementation:**
   Upon identifying the potential 10 features (per person), these features were incorporated on a particular company's stock to better study the information conveyed and gain more domain knowledge.

   One example of such refined features is: Linear Regression Forecast.



Next Steps:

Having narrowed down features for the clustering process, the next step is perform PCA on different stocks using these features to see if we can find clusters. It is also important to interpret the domain significance of the formed clusters. For instance, very clean (discrete) clusters can be formed considering the volume chart, but they are of little use if not validated further using some price or return indicator. With significant progress made in domain understanding, we have a strong base to dive deeper into feature engineering.

The final step would be portfolio diversification. Two major factors generally linked to a portfolio's fitness are expected returns and associated risk. To ensure that stocks with high correlations do not make up the entire portfolio, logic dictates that we maximize intra-cluster correlations and minimize inter-cluster correlations. 'Correlations' here, in turn, will be PC axes obtained after PCA. We test the robustness of the created portfolio using Markovitz's mean-variance portfolio theory (modern portfolio theory), which basically maximizes the asset returns for a given level of risk.

Link to the feature studies performed:
https://drive.google.com/file/d/1gQF84druyTsRTsv9BcEXGuzaKjpAT0_U/view?usp=sharing