SOEN 471
Big Data Analytics

Stock Market Analysis of the Tech Industry during the Pandemic

Project Summary

Team Members

Vashisht Marhwal: 40158888
Amro Lolooh:
Ross Giagkoudis: 40031093
Matthew Alquisada-Mansoori: 40064122

Project Summary

Motivation

With the rise of Covid-19 affecting the global economy, companies are faced with uncertainty and stocks are more volatile than ever. Various businesses are steadily adapting to this new era while others fall apart. But what differentiates these companies - why do some succeed and some don't? To start off, based on the collected dataset, our goal is to determine which few companies have taken advantage of the global pandemic and thrived in the economy. To continue, among the successful businesses, our next step would be to gather analogous information on their data in order to form a proposition determining why these certain companies were able to do well. With this information, we would be able to understand why and how the following firms achieved success and why their stock prices soared. This will enable us to direct other companies to follow the same footsteps in order to flip their financial state on the right side.

Algorithm

The algorithms used in this project are twofold: The first part of the project includes analyzing a stock market dataset to find companies that qualify as having "good financial performance" during the span of the past 2 years. This will be done with K-means clustering. Various factors such as stock prices, etc. will be used to filter out the good performers. This is an unsupervised learning technique since no target labels are present.Our clusters will be based on the performance in the last couple of years..

After that, multiple performance metrics and parameters are analyzed to find common patterns that will signify what are the determinant factors that result in good performance. For this, Hierarchical clustering will be used here.This would allow us to find data similarities in our performance cluster from Step 1.

Evaluation Metrics

Since we don't have the ground truth labels, we will use Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index to evaluate our clustering.

Data gathering and Libraries used

A large stock market price dataset will be used showing historic prices across the desired timeframe. The Dataset is "Huge Stock Market Dataset"[1] from Kaggle. Additionally, an API by a startup company Alpaca[2] will allow us to fetch prices of stock. Certain parameters will be analyzed depending on the nature of the filtered companies, that can include but are not limited to: number of users, engagement, financial activity.
The libraries that we will be using are Numpy, Pandas, Seaborn, Sci-Kit Learn and Apache Spark.

References

[1] https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

[2] https://alpaca.markets/