ChE 4320 Project – Group 5

Machine learning is an extremely useful approach to data analysis. This method involves analyzing patterns and trends in a given data set to understand the relation between the data. Further, specific machine learning techniques allow for the prediction of future behavior from the data set. Understanding machine learning can become valuable in the extremely data driven chemical engineering industry.

For the purpose of the project our group has chosen a dataset containing daily stock value data for ten corporations. The data spans from 2015 through 2025. Market data is a relatively straightforward report, making it a useful example set for machine learning. In the set, there are just under 2800 reported values for each company across the time range. Having a large pool of data will make it easier to analyze and predict how stock values behave.

The features for this dataset, the individual companies, report the same type of values. We can look at the trends for each respective feature, whilst also comparing the results from multiple features. Additionally, stock values are already analyzed and reported on. The context for changes in pace or extraneous data can be likely be found, as each of the companies is large and publicly traded.

Stock data is continuously reported, and trends are easily accessible with readily available analysis. For unsupervised learning this essentially serves validation of the analysis done for the purpose of the project. As stated earlier, the data is fairly simple, though we can use the unsupervised learning to compare the trends of each feature. Using a clustering technique may show companies that had similar stock behavior across the timeframe, creating potential for inference-making based on industry and world events. Similarly, dimensionality reduction will be useful in further simplifying the dataset to generalize the overall market behavior across the range. Again, this can be interpreted in the context of relevant global events.

The amount of data included will be useful for the supervised learning portion, as the sizeable trend history will contribute to a more accurate behavior prediction. Though supervised learning has not been discussed as much in class thus far, we are beginning to plan the approach to part three of the assignment. The expectation is that the technique used will base the future behavior for each feature on the general trajectory across the obtained data. We are curious to see how we are able to make a prediction with the natural rise and fall of stock prices.

The preliminary data workup within VS code has consisted of observing the distribution of stock price for each feature, which has revealed a reasonable positive skew with less weight on higher values. Due to varying respective range of stock prices for each feature, we felt a normalization would be appropriate to observe price trends beyond the base magnitude. The data seems pretty easily workable, and we are fairly confident it will serve as a good set for attempting machine learning.