**CNN Model for Stock Market Prediction**

**Data Analysis**

The data used for this project consists of 82 features from 5 markets: S&P 500 index, NASDAQ Composite, Dow Jones Industrial Average, NYSE Composite, and RUSSELL 2000. These features can be broadly classified into eight categories: primitive variables (close price and day of the week), technical indicators (technical features extracted from historical data like moving average), world stock markets (effect of other countries’ stock markets), exchange rate of US dollar in major countries, commodities (price of gold, wheat etc.), Big U.S. companies (stock prices of major companies like Apple), future contracts (effect of future prices of stocks) and other useful variables (eg: treasury bill rates).

Technical Indicators and others are often used for prediction of stock market fluctuations. However, sometimes the features cannot be hand engineered even by the most qualified finance experts. Therefore, CNNs are a great way to automatically extract such features which are relevant in predicting stock market fluctuations.

I have used the Tensorflow’s keras API in this project. Before using a ML or DL technique, it is necessary to pre-process the data. The data (consisting of 5 .csv files) were downloaded. It was read into a dictionary with the dictionary keys equal to the ‘Name’ column in the .csv files. Hence, the dictionary ‘data’ consists of 5 key value pairs, with each key being the name of the stock market and the value being a dataframe consisting of 82 features.

A ‘Target’ column is added in each dataframe such that if the percentage change in the Close Price with respect to the next day is positive, then the ‘Target’ variable for that day is set to 1, otherwise 0. The ranges of the features vary, as can be seen in the code. Hence, it is necessary to perform normalization. Normalization helps in keeping the variables in a certain range which helps during the training process.

Exploratory data analysis was also performed by a heatmap of the correlation matrix between the features. The heatmap showed that almost none of the variables had a significantly high correlation with the target. This signifies that the target is really random and hence, these features alone may not be enough to directly predict stock market fluctuations. Hence, it is important to use automatic feature extraction tools like CNNs which can extract important feature information automatically.

It is to be noted, that common feature engineering techniques like differencing and logs maybe used. However, because the choice of features are extensive and covers a whole lot of technical indicators like Relative Change of Volume, Rate of Change, Exponential Moving Average, I think they will act as better predictors for our model.

Following data analysis and engineering, it is now possible to build the model.