Stock Price Forecasting

I have chosen, for my capstone, a dataset from kaggle that contains information from several Stocks over the past 10 or so years. The information included in each stock is for each day, and the columns are as follows; Date, Open, High, Low, Close, Volume and Name of stock. I have chosen to use the Stock information from AAPL, AMZN, GOOGL, NKE, and VZ.

1. **Problem Identification**

The stock market plays a vital role in our economy. It is often used as a sentiment to how the economy is performing, and impacts the GDP of our country. For years, industries, companies, and individuals have looked to capitalize on the economy to help grow their finances, regardless of whether they are trying to grow funding for their business, retirement account, or simply trying to get a little extra 'fun money'. Being able to accurately predict future prices and trends based upon the previous entries can be seen as beneficial to anyone. For this particular problem, I would imagine that trend forecasting would be more important to most than accurately forecasting prices.

2. **Data Wrangling**

Since this is a dataset from Kaggle, obtaining the csv's of the stocks will not be difficult. From there I will go through the data that is provided to ensure that the data is uniform. I will make sure that the variables make sense and are populated with the right data types and values. I will make sure that there are no missing values or errors in entries (anything that is obviously different from the others). Since this is historical stock data, any missing or incorrect values can be directly imputed with the correct ones. Since I am dealing with forecasting models, I will need to either convert the date into a datetime object or create a new datetime object and drop the date column.

3. **Exploratory Data Analysis**

During this step, I will still be looking at the integrity of the data, but I will be looking more for trends or identifying relationships between the features. I know that I will only include two columns into my models, the date, and the feature I choose to represent the price. I am assuming I will choose either Open or Close, as the Volume, High, and Low values, while important, do not directly correlate to trends within the pricing of the stock as well as the Open and Close does. I ultimately chose to use the Close values. Initial findings show that the last week of the year proves to be the most volatile as far as price fluctuations are concerned.

4. Pre-Processing and Training Data Development

During this step, I will be actually looking at how to impute any missing data. If any of the metrics did contain different units of measure I would want to transform them all to match at this point. For this data, if there are any missing values I will simply impute the correct information since all historical stock information is available. There are no categorical data columns containing strings, so I will not need to create a new column assigning the value of a number to a given category. I will also use cross-validation tools and model selection to make sure I accurately check the model performance.

5. Modeling

Here I will be working through the different models to determine what set has the highest level of accuracy. In this case I will be looking at the RMSE of the different models and comparing how each type of forecasting produced different RMSE values. The models that I will be using are FBProphet, ARIMA, and Exponential Smoothing. Because I am using information from 5 stocks, I will create a total of 15 models and compare the RMSE of each.

6. Documentation

During this final stage I will be taking my models and findings and summarizing them into one report. I will take my suggestions for utilizing any of the models created to forecast stock prices/trends and present them to the audience, while then going back and explaining in everyday terms what I was doing to get to that ultimate result. I will not highlight every single step in the project, moreso the thought processes behind a specific step.