**Abstract:**

Stock market predictions is a highly sort after field. The volatility and uncertainty of this field makes it very attractive to researchers. There are many papers published on this topic but the field is so vast and the prediction capabilities of these papers is so limited that there is still a considerable expanse to explore. I will try to implement a more tried and tested approach that is to use a trend based predictions. In this paper, I will discuss some of the strengths and weaknesses of such an approach. More specifically, I will implement a machine learning approach to a trend based stock prices data using spark. I will use the same approach without spark integration to test the difference in running time, I will collect the accuracy of the predictions as well.

**Introduction:**

Stock market was introduced in the 1790’s. With every passing year the amount of capital in trading is increasing, the number of major companies in the market is ever increasing and the people involved is also increasing. This creates many variables that have an impact on the stock prices of a company. Some of the variables have more impact than others, a more accurate prediction could be achieved if most of these variables are considered for the predictions. This means that a lot of data is required and all this data must be learned quickly so as to predict the values. If the software takes hours or days to make a prediction, then by the time the prediction is made the deadline may have passed and we may have more data to sort through. Speed as well as accuracy is important in this case. Using Big Data techniques such as spark can help with this problem. Broadly speaking, there are 2 main techniques for predict the stock market prices - Trends based and Event based. Trends include the price of stock at a given time, Event based includes the impact of certain event on the stock prices. Using an event based approach in this project will be out of scope because of the time constraints. For this project I will use just the trends. For most traders, trends are the most important criteria. A lot of them are not trend setters, they just follow the existing trends. Logic dictates that if we can understand the trends then we can predict the approximate stock price of a company at any given time.

In this project, I will be using this approach to predict the stock prices. This is not the most accurate or the most profitable approach but it is a good starting point for most traders. I will use neural networks via TensorFlow to make the predictions and spark to decrease the run time of the software.

**Strategy used:**

The best method to make predictions using data is to apply some machine learning technique. Since this a time series data any regression method would help for this application I will use neural networks. I would like to use TensorFlow for this project.

TensorFlow is an open source library created by google, it’s flexible architecture allows us to compute complex software over one or more CPUs and GPUs. There are methods to integrate TensorFlow with spark. GPUs can speedup a software by up to 10x on a local configuration but an integration with spark will come into its own when the data is so large that it can’t fit into one’s machine memory or the machine’s GPU memory. Ideally a TensorFlow software working GPU with spark should be the fastest approach if we have an extremely large amount of data. For this project however, since I am working on relatively smaller data I don’t anticipate the need for such power. I intend to use CPU with spark integration.

For this project, I plan on predicting the stock prices of a certain company. I haven’t decided which company yet but it’s not really important at this juncture. The data can be obtained from Yahoo Finance.

**Functions:**

As mentioned earlier, this software will be built to try and predict the stock price of a company. It will be more useful to an average trader rather than a finance firm.

The input will be a dataset consisting of the open, close, highs, lows and volume of stock prices of a company over a certain period. The data is available for over some decades but all the data is probably not required. Studies have shown that the stock prices are more dependent on recent past rather than over many years. The best time period will be decided based on the time consumed and the accuracy of the final result by trial and error.

**Testing:**

For any machine learning application, a train data and a test data is required. For this application I will use about 80% of the data for training and the remaining 20% for testing. The test data will be the latest 20% from the most recent history as this will be the closest data to the present and hence most useful to understand the quality of the software.

**Evaluation:**

For this kind of application, speed as well as accuracy is important. To evaluate this, I will use the same application without the spark integration to test the difference in the run time of the software.

The second part of the project is to test the accuracy of the predictions, there is no real way to compare the accuracy obtained from the software. A lot of the papers in this topic seems to show the results but don’t really compare their result to some other papers. Quite a few papers use the software as an agent and compare the profit made by the agent with some monkey agent making random choices. This approach doesn’t really prove anything. Since I can’t really compare the performance of the software, I am targeting an accuracy of at-least 70% with a 0.5% error in the value. This target may not seem very high but given the limited number of variables used and the volatility of the market, this is the best accuracy one can hope to achieve. This is also the reason why stock market is such an intriguing field for so many researchers, getting an accuracy of over 90% is virtually impossible. However, for an average trader an accuracy of over 50%, probably means a profit on their investment.

**Project Timeline:**

**Bibliography:**

1. Paul D. Yoo, Maria H. Kim, Tony Jan, “Financial Forecasting: Advanced Machine Learning Techniques in Stock Market Analysis”, *9th International Multitopic Conference, IEEE INMIC 2005*, 19 March 2007.
2. Girija V Attigeri , Manohara Pai M M, Radhika M Pai, Aparna Nayak, “Stock Market Prediction: A Big Data Approach”, [*TENCON 2015 - 2015 IEEE Region 10 Conference*](http://ieeexplore.ieee.org.ezproxy2.library.colostate.edu:2048/xpl/mostRecentIssue.jsp?punumber=7365563), 07 January 2016.
3. Radu Iacomin, Stock Market Prediction, *“System Theory, Control and Computing (ICSTCC), 2015 19th International Conference”*, 09 November 2015.
4. Kevin Jhaveri, Dhiral Shah, Saloni Bhanushali and Era Johri, Financial Market Prediction Using Hybridized Neural Approach, *“*Computation of Power, Energy Information and Commuincation (ICCPEIC), 2016 International Conference*”*, 01 September 2016.
5. Jeff Dean, Monga, Rajat, et al, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, *“Google Research”*, 09 November 2015.