CSCI 4502 Project Proposal

Chandler Baggett
chandler.baggett@colorado.edu

Dmytro Ryzhkov
Dmytro.Ryzhkov@colorado.edu

Jacob Toomey
Jacob.Toomey@colorado.edu

ABSTRACT
In this paper, we describe our proposal for a project concerning mining the stock market for data in order to search for discernible statistical trends.

General Terms
Finance, Economics, Stocks, Stock Market, Trends, Prediction, Prices, Patterns, Future Performance, NASDAQ

Keywords
Machine Learning, Python, Data Mining

1. PROBLEM STATEMENT/MOTIVATION

In our project, we plan to analyze past data of stock market fluctuations in order to attempt at making an accurate prediction of future stock prices. Our motivation behind conducting this research is to better understand data mining and statistical trends using real-world data. We intend to answer the following questions:

- Is it possible to find patterns in previous stock prices and the history of the stock market in order to predict its future performance?
- Can we utilize data mining and machine learning to recognize certain patterns in certain stocks that can predict how a stock might behave next?

2. PREVIOUS WORK

- Genetic Algorithms have been used to try and predict the stock market.
- Banks and High Frequency trading firms hire analysts to predict the stock market and have talented developers writing algorithms and utilizing machine learning to try and predict the stock market.
- There are many papers written on the subject and a lot of academic research going into utilizing machine learning in order to predict the stock market.

2.1 Application of data mining techniques in stock markets: A survey by Ehsan Hajizadeh, Hamed Davari Ardakani, and Jamal Shahrabi

2.2 Financial Stock Market Forecast using Data Mining Techniques by K.

Senthamarai Kannan, P. Sailapathi Sekar, M.Mohamed Sathik and P. Arumugam

2.3 Stock market time series forecasting with data mining methods by Milan Csaba Badics

3. PROPOSED WORK

Our goal is to feed preexisting datasets that contain stock prices over a period of time into machine learning algorithms in order to make more educated predictions about the future prices of individual stocks. Our methodology will include code written in Python and consistent usage of the Scikit-Learn Library found in Python. After we have used Machine Learning tools to learn from past data, we will then interpolate our results to predict what the future price of a stock may be. Our goal is to make a hypothetical "profit" based on our statistical findings by running our algorithms on a past data set (or interpolated predictions on real-time day to day stock prices) and checking to see if there is any correlation between how accurate our predictions are of different stocks.

4. DATA SET

There are many data sets containing information on stocks (both individual and total stock market) online.

and http://www.nasdaq.com/quotes/ will both serve as good sources as they have clean stock price information in .CSV files.

5. EVALUATION METHODS

In order to evaluate how efficient our machine learning tool is, we will compare it with results that other people have received from their machine learning algorithms and past data sets of stocks. There are several sources online that show returns that investors have from utilizing algorithmic trading in the stock market as well as past data sets of stock performance.

6. TOOLS

We will be using Python in order to write code that will use machine learning in order to make predictions. We will be using the Scikit-Learn library in Python in order to help us make calculations.

7. MILESTONES

1) Data sets have been downloaded.
2) Python Library Scikit-Learn has been researched and is decidedly our main tool for conducting machine learning and data mining of the stock market.
3) We have met and decided to test our machine learning algorithms on the top 50 stocks as of Friday March 3rd, 2017 listed by Company Name (Symbol): Apple (AAPL), Alphabet Inc. (GOOG), Microsoft (MSFT), Berkshire Hathaway (BRK-A), Amazon (AMZN),

Facebook (FB), Exxon Mobil (XOM), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), Wells Fargo (WFC), General Electric (GE), AT&T (T), Bank of America (BAC), Procter & Gamble (PG), Wal-Mart (WMT), Chevron (CVX), Visa (V), Pfizer (PFE), Verizon Communications (VZ), Merck (MRK), Coca-Cola (KO), Comcast (CMCSA), Home Depot (HD), Oracle (ORCL), Disney (DIS), Cisco Systems (CSCO), Intel (INTC), IBM (IBM), Philip Morris International (PM), Citigroup (C), PepsiCo (PEP), UnitedHealth Group (UNH), Altria (MO), Amgen (AMGN), Mastercard Inc. (MA), Schlumberger (SLB), Kraft Heinz Co (KHC), 3M (MMM), MedTronic (MDT), Boeing (BA), McDonald's (MCD), Goldman Sachs (GS), Charter Communications, Inc. (CHTR), AbbVie Inc. (ABBV), Celgene (CELG), Bristol-Myers Squibb (BMY), Honeywell (HON), NIKE Inc. (NKE), Walgreens Boots Alliance (WBA), U.S. Bancorp (USB).

8. SUMMARY OF PEER REVIEW SESSION

Following our presentation, we were faced with the possible issue of gathering enough data points in order to make the most educated prediction of future stock prices. In order to address this concern, we will be running our algorithm on the top 50 stocks (S&P).

Another issue that was considered was different levels of volatility in different sectors: does energy behave the same as health care? What about real estate? This is something that we will explore further in our research.