Data Science Final Project Overview

Dan Raab
December 21, 2017
draab678@gmail.com

Contents

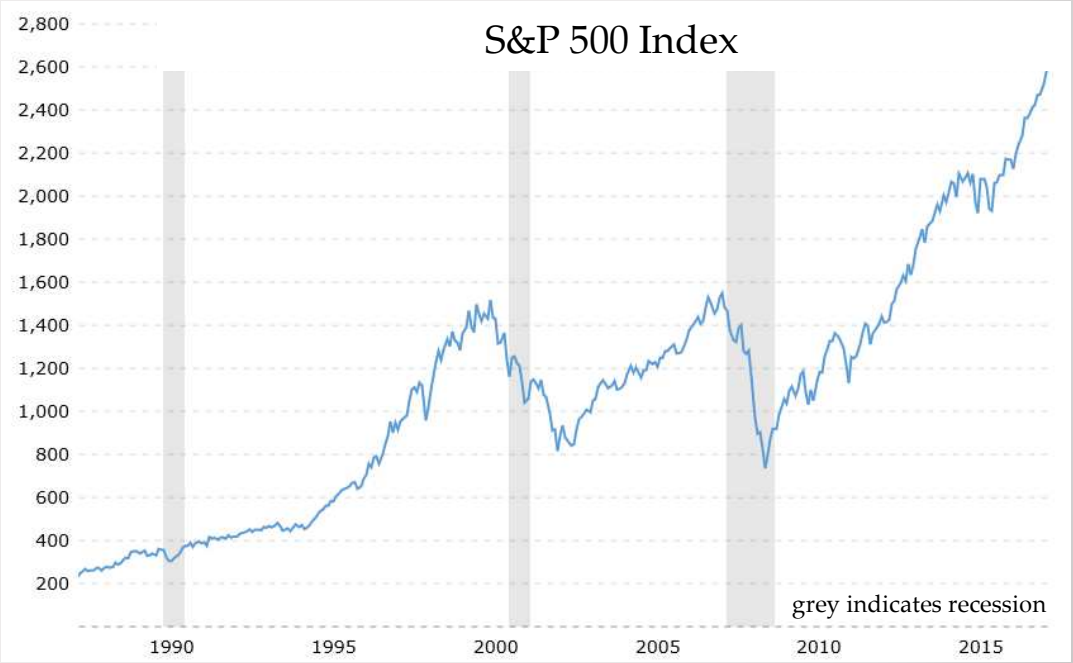
- Project Problem and Hypothesis
- Datasets
- Domain Knowledge
- Project Concerns
- Outcomes

Project Problem and Hypothesis

- The stock market is one of the most common places to invest for long term asset growth
- However, it is constantly being impacted by short term factors which cause tremendous volatility
- This volatility can cause dramatic fluctuations in the value of investments
- These dramatic fluctuations will cause negative impacts to risk adjusted returns
- Investors, including professional portfolio managers, may find it difficult to interpret the constant stream of news, and to assess the impact this news will have on the stock market
- Attempting to do so may be counterproductive and time consuming
- My hypothesis is that I can create a machine learning algorithm that will read news data, and detect patterns within this data which will provide signals to buy and sell stocks
- This algorithm should provide a portfolio optimization tool which may provide enhanced risk adjusted results compared to a non-optimized exposure

Stock Market Performance

- The average annualized total return for the S&P 500 index over the past 90 years is 9.8 percent [1]
- However, the returns do not come in steadily each year, and the market will typically go through dramatic swings in both the positive and negative directions
- For example, "in the 89 years from 1928 to 2016, only six finished with a gain in that range that we think of as a 'typical' annual return" [1]
- As the chart illustrates, while the S&P 500 has gained over 1000% over the last 30 years, it also experienced 2 periods during which it lost around half its value
- Even the most steadfast investor may find it difficult to stay in the markets when their investments periodically decline by that much



Source: http://www.macrotrends.net/2324/sp-500-historical-chart-data

1 https://www.cnbc.com/2017/06/18/the-sp-500-has-already-met-its-average-return-for-a-full-year.html

Datasets

- My goal is to use machine learning to read news and interpret the news as positive or negative for the stock market
- The prediction may be binary, ie Buy or Sell, or may be extended to weighted signals, for example Buy 100%, Buy 50%, Buy 25%, Hold, Sell 25%, Sell 50%, Sell 100%, depending on the results
- The datasets that exist include for example : Combined_News_DJIA.csv, available for download on Kaggle
- The provider of the data described it as follows:[1]

  1) News data: I crawled historical news headlines from Reddit WorldNews Channel (/r/worldnews). They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date. (Range: 2008-06-08 to 2016-07-01)

  2) Stock data: Dow Jones Industrial Average (DJIA) is used to "prove the concept". (Range: 2008-08-08 to 2016-07-01)
- Another historical news data set available on Kaggle includes abcnews-date-text.csv, which is described as follows: This contains data of news headlines published over a period of 14 years. From the reputable Australian news source ABC (Australian Broadcasting Corp.) [2]
- I would plan to train/test/validate on one dataset, and further test it using alternative news data sets along with alternative stock market data sets, for example testing the Australian news versus the Australian stock market
- Ideally the algorithm will work well across multiple countries and stock markets
- In order to be interpreted the dataset will need to be broken down into words and/or phrases, and this data may further be broken down into categories of words and/or phrases, such as “positive” or “negative”

1 https://www.kaggle.com/aaron7sun/stocknews

Sample of News Data

Following is a sample of Combined_News_DJIA.csv

Label	8/14/2008
Top1	b'All the experts admit that we should legalise drugs '
Top2	b'War in South Osetia - 89 pictures made by a Russian soldier.'
Top3	b'Swedish wrestler Ara Abrahamian throws away medal in Olympic hissy fit '
Top4	b'Russia exaggerated the death toll in South Ossetia. Now only 44 were originally killed compared to 2,000.'
Top5	b'Missile That Killed 9 Inside Pakistan May Have Been Launched by the CIA'
Top6	b"Rushdie Condemns Random House's Refusal to Publish Novel for Fear of Muslim Retaliation"
Top7	b'Poland and US agree to missle defense deal. Interesting timing!'
Top8	b'Will the Russians conquer Tblisi? Bet on it, no seriously you can BET on it'
Top9	b'Russia exaggerating South Ossetian death toll, says human rights group'
Top10	b' Musharraf expected to resign rather than face impeachment'

Sample of Stock Market Price Data

Following is a sample of DJIA_table.csv

Date	Open	High	Low	Close	Volume	Adj Close
07/01/16	17924.24023	18002.38086	17916.91016	17949.36914	82160000	17949.36914
06/30/16	17712.75977	17930.60938	17711.80078	17929.99023	133030000	17929.99023
06/29/16	17456.01953	17704.50977	17456.01953	17694.67969	106380000	17694.67969
06/28/16	17190.50977	17409.7207	17190.50977	17409.7207	112190000	17409.7207
06/27/16	17355.21094	17355.21094	17063.08008	17140.24023	138740000	17140.24023
06/24/16	17946.63086	17946.63086	17356.33984	17400.75	239000000	17400.75
06/23/16	17844.10938	18011.07031	17844.10938	18011.07031	98070000	18011.07031
06/22/16	17832.66992	17920.16016	17770.35938	17780.83008	89440000	17780.83008
06/21/16	17827.33008	17877.83984	17799.80078	17829.73047	85130000	17829.73047
06/20/16	17736.86914	17946.35938	17736.86914	17804.86914	99380000	17804.86914

Domain Knowledge

- I have worked in the financial markets for over 20 years, with a particular focus on developing financial indexes
- As part of this work I have developed trading models with trading decisions driven by price data and economic data
- Given my participation in the markets I have constantly read news and interpreted it to make decisions, both in terms of trading as well as model design
- This approach to analyzing natural language for trading purposes has been developed by trading firms and has been experimented with by academics…work is ongoing
- I believe this area has not been fully developed either with respect to financial market predictions or other business applications: the continuous evolution of data, technology and financial markets should create ongoing opportunities for development
- "The real challenge for enterprises is getting value from a sea of social media posts, images, email, text messages, audio files, Word documents, PDFs and other sources that make up the other 80 percent of data that can't be understood by computers—information otherwise known as unstructured data" [1]
- Once the algorithm is successful in one market and region, the goal would be to expand it to other markets and regions, such as bonds and commodities, including both regional and global news and markets

1 https://www.ibm.com/blogs/watson/2016/06/natural-language-processing-transforming-financial-industry-2/

Project Concerns

- One area of concern is in terms of accuracy: How well will the algorithm predict up and down movements of the markets
- In addition, there are many choices of news data, and many choices of market indicators, ranging from broad market indexes, to sector indexes, down to individual stocks
- A concern will be to define the data set either too narrowly or too broadly, and thereby miss valuable predictive relationship which may exist given another subset or larger set of the data
- Another concern is if the algorithm is successful on a limited data set will it be more universally applicable
- Other concerns include the timing of the news and how long its predictive value exists: for example, is a one day lag too much, as in this experiment all data is available as of the end of the day, when in reality the news was released during the day. Will too much time have elapsed by the time the next day's market movements occur?
- I will test same day predictive value, but the real use will be for forward looking predictive value, where there will be time to implement actual positions in the markets

Outcomes

- The output will be a vector of predictors, possibly +1, 0, and -1, or potentially additional values within that range
- These predictors will indicate that the stock market will go up, will not change, or will go down
- If these predictors are accurate, then they could be used as trading signals, for example to buy, hold, or sell on the day after each signal is calculated
- The final test will be to calculate a theoretical portfolio based on these buy/hold/sell signals and tradable market prices
- The model will need to be somewhat complex, given that this space that has so much attention attracted to it: I expect that simple models will likely not be successful or they would already have been created and the value they presented arbitraged away
- Anything above 50% successful predictive capability will be a good outcome, however I want to see 60% or more to make sure the model provides enough of an advantage to make it useful and cover any expected trading costs and slippage in actually executing it in the market
- A real value of this project will be if it is accurate over multiple data sets, spanning multiple markets and geographic regions. This would indicate that it is robust and not too heavily fit to one particular market, region, or time period
- If that is the outcome then I would expect it to continue to perform in the future and to be a valuable tool for investment optimization