capstone_proposal

December 18, 2018

1 Machine Learning Engineer Nanodegree

1.1 Capstone Proposal

Chieh-Ling Hsieh
December 17th, 2018

1.2 Macro economic indicators for stock market and investment strategy

1.2.1 Domain Background

Equity market has long been an interesting subject of study for academic, investors, market makers, but despite the amount of studies, people still never able to reach a conclusion on this subject with its ever shifting behavior. From Nobel price winner Eugene Fama's factors modeling, to Nobel price winner Robert Shiller Cyclically adjusted price-to-earnings ratio (CAPE) and behavior economic theory, to Bridgewater Associate Ray Dalio Debt Cycle, there are various theory and factors that are some way of observing the market. While there are also theory that all the prediction are worthless as suggested Burton Malkiel in his famous book of A Random Walk Down Wall Street. Retail stock investment expert Andre Kostolany has his own theory of performance = money supply + sentiment.

One of the possible reason for the various conflicting theories to co-exist are the market are co-decided by all the players in the market, where the human decision might not follow strict pre-defined theorem. Most of the studies also suffer from lack of data during the time of studies as data collection were not as easy and analysis tools were not as pervasive. With more abundant data nowdaay, there could be potential to perform some innovated studies and could provide new insights into this problem.

1.2.2 Problem Statement

While the global economic systems are enormously complex, there have been either more formal approaches or informal rule of thumb experience for certain big indicators that could move the market more significantly than others. Data includes economic statistics and leading indicators, long term interest rate, or even the price action of the market itself could all be used to as in hypotheis that are factors that could affect the market.

To formulate a trading strategy, we can formulate the decision point to be very simple. For example, given the known states of the features we gather, do we expect the market to go up, or go down? Then expectation can be translated into buy and sell decision. In machine learning terminology, this is surprvised learning classification problem. We could set a look ahead to be exepcted result. For example, will market go up or down after one month from now. Given that we 'known' the future price movement, we can use that as label for the training in training data set, and use it to test on future data set.

1.2.3 Datasets and Inputs

The study is interdisciplinary in nature that we want to be free flowing to not limiting the consideration of features to traditional ones. Also, the data are maintenaced, published and made available by different organizations and in different formats.

For financial market data and some of the macro economic data, we plan to gather the data source from

- Yahoo Finance

Specifically we will be looking for ˆGSPC, which is S&P 500 index. This is broad market based index and will give good overall stock market performance. ˆGPSC is a single numeric value. We would also get interest rate data as well, which is a single numeric value of percentage.

For economic index data, we plan to gather from

- Conference Board We can obtain leading economic indicator. The indictor itself is a numberic value. However, also publish is monthly change percentage. Conference Board change the indicator value occasionally so monthly change value could be more reliable over longer period of time.

For market sentiment data, we plan to gather from

- Chicago Board Options Exchange Specially, we can look for Put/Call Ratio, which is a numeric value of ratio of outstanding put and call.

There are more similar macro economic or sentiment time series data can be obtained and experiment with. And we plan to add a few more through the process.

1.2.4 Solution Statement

After gathering the data, we will then attempt to use the past to predict the future. We can use features at a given snapshot of time, to predict the price move of one month after. By setting the prediction to be some time after like one month or 30 trading days, we could also use the price itself as part of the features because the setting avoid look ahead bias because price used as feature is always one month before the price used in labeling of up or down market. We could also not including price as feature but one month setting fit the problem requirement very well (want to predict the furture) and the settings allow the freedom to include everything gather as feature if we desired.

1.2.5 Benchmark Model

Depend on the exploration process, and the models chosen to better fit and predict the data, the metric builtin with the model that are more suitable for the specific method will be used to gather metric to measure fitness score of the model.

Once a model is found, we will try to build a decision model to suggest the trading stragety of the market. The effectiveness of the strategy can be benchmarked agasinst commonly known investment strategy, like passive investment, which is the stock index itself.

1.2.6 Evaluation Metrics

Once we have a chosen strategy and benchmark, evalulation of them tested such as * accuracy score * precision score * recall score * f1 score

It has been proved that in investment, people would rather high likelihood of not losing money rather that low likeihood of earning a lot more money. In our research, we plan to classifiy the features either as bull or bear market indicator and label bull market as 1. In the assumption that people are not shorting the market, they will be more happy on high accurarcy of bull market call, as this is make them not losing money. Hence we will choose accuracy score as our metric.

1.2.7 Project Design

The project will be consist of several phases

Data collection, cleaning, and wrangling The data will be collected from various locations, and go through cleaning and wrangling process to ensure they are can be worked together. For example, time duration, time frequency of each data source might not be the same, and the range of range might not be directly compariable and require normalization.

Model fitting and validation Several models of supervised learning methods such as decision tree, SVM, Naive Bayes, could be used to fit the model and predict the outcome. We can set aside certain period of time period to be used for training. And certain period of time to be used for testing. Time series data are continuous so the data set has to remain so when splitting them for training.

Stragety development and benchmarking The insight we obtained from the development and model, should lead us to delveopment of a trading model of investment. The model can then be backtested against common investment strategy.

Reflection The findings and resutls of each phases will be assessment for value to ordinary investors and future area of study.

```
In [ ]:
```