Machine Learning Engineer Nanodegree

Capstone Project

Chieh-Ling Hsieh

December 21st, 2018

Macro economic indicators for stock market and investment strategy

I. Definition

Project Overview

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. The importance of this subject has become more and more important as development and policy progress. More and more governments shifting the retirement from entitlement plan to private based investment. For example, 401K retirement plan in United States has and will be increasingly become a major income source for people entering retiring age. The importance not only has significant in overall health of the economic, but also matter to everyone's well-being in the increasingly more capitalized society.

From Nobel price winner Eugene Fama's factors modeling [1], [2], to Nobel price winner Robert Shiller Cyclically adjusted price-to-earnings ratio (CAPE) and behavior economic theory [3], to Bridgewater Associate Ray Dalio Debt Cycle [4], 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[5]. Retail stock investment expert Andre Kostolany has his own theory of performance = money supply + sentiment.[6]

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 strategy based on tested and proved 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 abundantly available data today, there could be potential to perform some studies and could provide new insights into this problem.

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 hypothesis that are factors that could affect the market. By looking at the correlation relationship between them, we hope to identify if any of them have significant, even in coincident, leading, o lagging way.

If the significance among these data are found to be affecting the market, a strategy can be developed to provide recommendation of investment, for as simple as whether to stay in the market when the current state suggests further increase of the market equity, or stay out of it when the current state suggest otherwise. The strategy can be benchmarked against common strategy such as the very popular passive investment, e.g., buy and hold the index fund. Or strategy of keeping balanced asset allocation between stock and bond.

The data we obtained are from various organizations. Although all of them are time series data, they will have their own format and different time frequency of the release of the data, daily, weekly, or monthly. The duration, that is, start and end date of the data could be different as well. We will collect that data, choose the convergence of time duration where all data are available, and merge the time frequency of all into daily frequency. This will create some discreteness of the data for data with longer release frequency, but this is best matched to the real world, without additional exploration or project which would introduce more variables.

We plan to use snapshot of features on a particular day, to predict the future price movement direction, going up or going down. But stating problem this way, it is a classification problem that we use the values of a set of features, to predict whether it is a bull or bear signal - a boolean value. This type of problem can be dealt with using supervised learning algorithms such as SVM, Naive Bayes, and Decision Tree.

Metrics

There are several scores available for evaluation of supervised learning algorithms, such as:

- accuracy score: TP+TN/TP+FP+FN+TN
- precision score: TP/TP+FP
- recall score: TP/TP+FN
- f1 score: 2*(Recall* Precision) / (Recall + Precision)

Where True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN) [7]

It has been proved that in investment, people would rather high likelihood of not losing money rather that low likelihood of earning a lot more money. In our research, we plan to classify the features either as bull or bear market indicator and label bull market as 1. If users use the model to do both long short trading strategy, they will need both bull and bear signal to be correct, hence accuracy score is important. In the assumption that people are not shorting the market, they will be more happy on high accuracy of bull market call, as this is make them not losing money. In such case, precision score is more important. We will use accuracy score as major benchmark and precision score as auxiliary benchmark.

II. Analysis

Data Exploration

The data we explored are maintained, published and made available by different organizations and in different formats. One unique characteristics of all the data is they are all time-series data. However, they might have different period of time that they collected the data, and have different frequency of data they are able to gather and collect.

For example, trading related data are available daily. While sentiment data based on survey are available weekly. Macro economic indicators and margin debt statistics are available only monthly. Some wrangling are necessary to be able to bring all the data to common format and background for processing.

Financial market data and some of the macro economic data, are gathered from

- Yahoo Finance

Economic indicator data, are gathered from

- Conference Board

Treasury interest rate data, are gathered from

- Federal Reserve Bank of St. Louis

Market sentiment data, are gathered from

- Chicago Board Options Exchange
- American Association of Individual Investors
- National Association of Active Investment Managers

Data on the use of margin trading, are gathered from

- Financial Industry Regulatory Authority, Inc.

The data collections process were done programmatically but are discussed in detail as the process are out of scope of this research.

Exploratory Visualization¶

Given all are time series data indexed by the date, we plot simple time series for those data.

Stock Price Data
The price data, S&P 500 Index has daily frequency value beween 1000 to 3000.

	Close
count	3261.000000
mean	1663.984090
std	532.328386
min	676.530029
25%	1277.930054
50%	1470.729980
75%	2067.889893
max	2930.750000

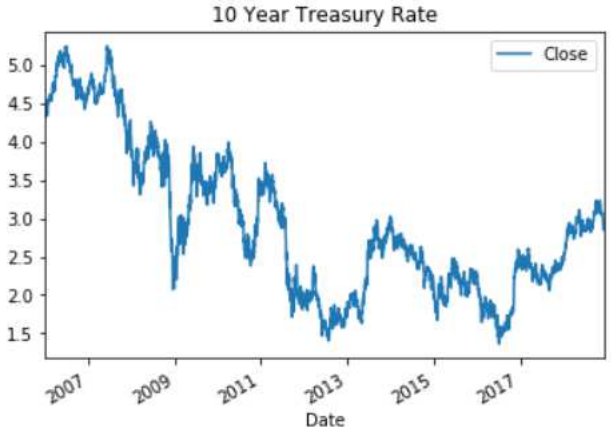
Date	Close
2018-12-10	2637.719971
2018-12-11	2636.780029
2018-12-12	2651.070068
2018-12-13	2650.540039
2018-12-14	2599.949951

10 year US Treasury Interest Rate

The 10 Year Treasury Interest Rate has daily frequency and rate between 0% to 6%.

	Close
count	3261
unique	1976
top	2.491000
freq	7

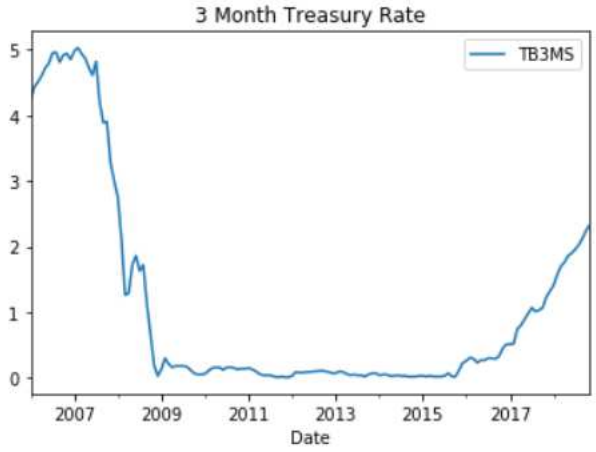
Date	Close
2018-12-10	2.856000
2018-12-11	2.879000
2018-12-12	2.906000
2018-12-13	2.911000
2018-12-14	2.891000



3-Month Treasury Bill: Secondary Market Rate

3 Month Treasury Interest Rate has monthly frequency and range between 0% to 6%.

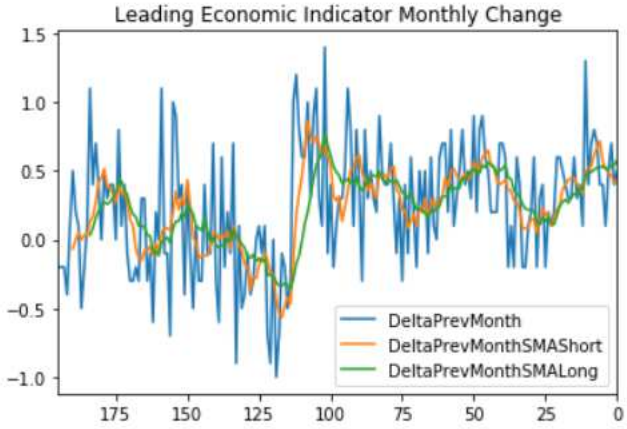
Date	TB3MS
2018-07-01	1.96
2018-08-01	2.03



Conference Board Leading Economic Indicator

Leading Economic Indicator's Monthly Percentage Change has monthly frequency and value between -1.0% to 1.5%. The data appear not well aligned on date, as other data set. As the "Date" is not used as index. When we bring all data set together this will be addressed.

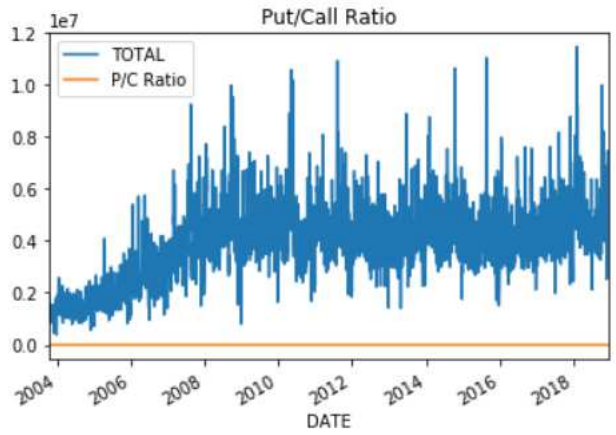
	DeltaPrevMonth	ReleaseDate	DeltaPrevMonthSMAShort	DeltaPrevMonthSMALong
4	0.1	2018-06-21	0.516667	0.516667
3	0.5	2018-07-19	0.483333	0.508333



CPOE Put/Call Ratio

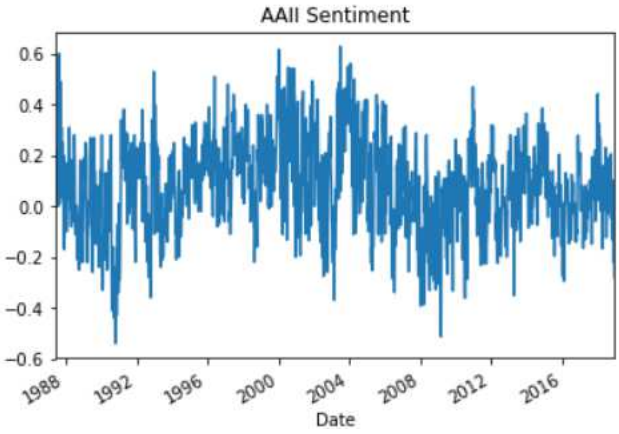
Put/Call Ratio has daily frequency and value between 0 to 1.2

	TOTAL	P/C Ratio
DATE		
2018-12-10	6229012	1.07
2018-12-11	4384177	0.97



American Association of Individual Investors Sentiments

Date	Bullish	Neutral	Bearish	BullMinusBear
2018-11-15	0.350877	0.289474	0.359649	-0.008772
2018-11-22	0.252525	0.276094	0.471380	-0.218855

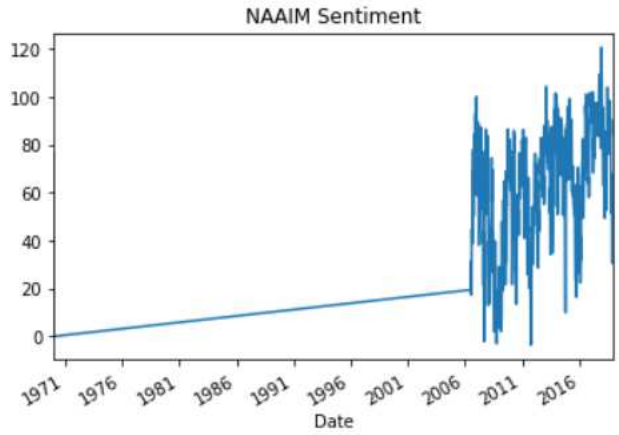


National Association of Active Investment Managers Sentiment

AAII Sentiment is survey data of investors. It has a weekly frequency. There are data for bullish, bearish, and neutral sentiment. To get a common value, we choose to use bullish percentage minus bearish percentage as the reference value for the overall sentiment.

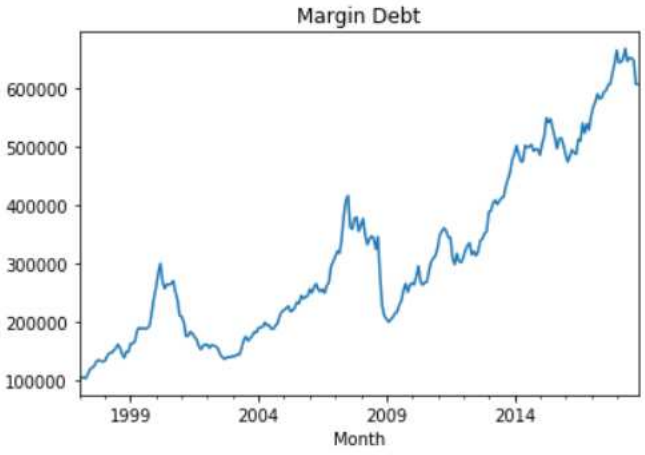
NAAIM Sentiment is survey data of investors. It has a weekly frequency. To get a common value, we choose to use mean as the reference value for the overall sentiment.

Date	Mean/Average	Most Bearish Response	Quart 1 (25% at/below)	Quart 2 (median)	Quart 3 (25% at/above)	Most Bullish Response	Standard Deviation	NAAIM Number	S&P 500
2018-11-14	35.13	-100	7.00	45.0	60.00	107	48.37	35.13	2722.18
2018-11-21	30.55	-200	2.50	30.0	79.00	130	73.89	30.55	2641.89



FINRA Margin Debt
Margin debt has monthly frequency.

Month	MarginDebt	PctChangesPrevMonth	MarketCap	MarginOfMarketCap
2018-07-01	652790.0	0.009057	3.418454e+13	0.019096
2018-08-01	652395.0	-0.000605	3.521908e+13	0.018524



Algorithms and Techniques

The data we gather can be grouped into following categories:

- Macro Economic Data
  - Leading Economic Indicator: various of collected economic data and appear to be able to predict the GDP in the near future.
  - US Treasury rate, including 10 year and 3 month duration.
- Investors Sentiment
  - From survey of investors organizations, American Association of Individual Investors, and National Association of Active Investment Managers

    - Put/Call Ratio, which are actual stock option traded that can be used as a measure of how many are betting on being a bull or a bear market.
- Equity price
    - The price of equity itself can be a reflection of how well economic is doing and how optimistic the investors are.

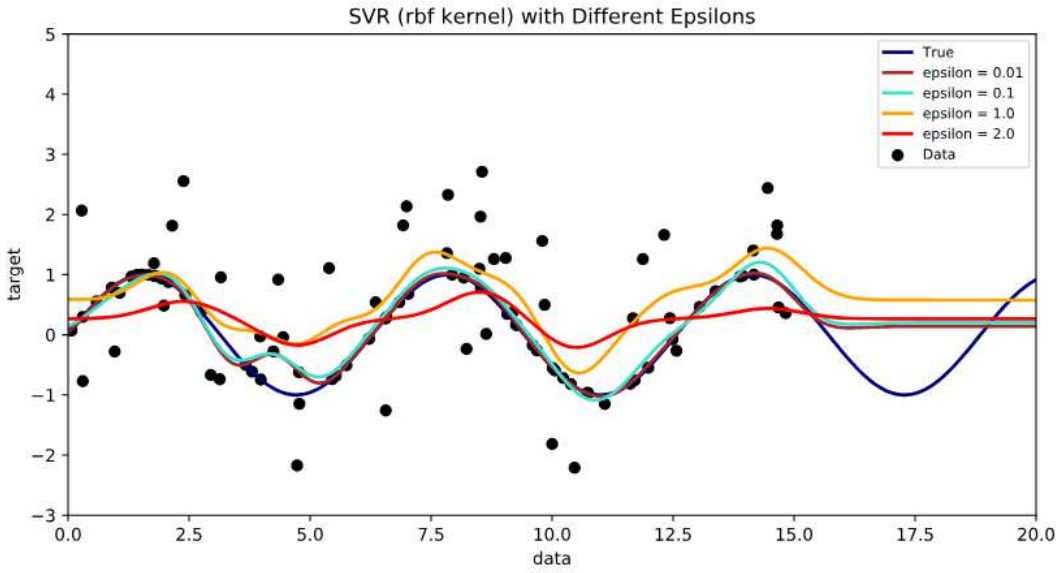
As the data are not of the same frequency, we will have to adjust the data so that they are of the same scale. We decided to put all data on the same daily frequency, and for weekly and monthly data, extend their value to the next release cycle. This will make the adjusted data appeared discrete instead of continuous. Although awkward, but more closely matched to real world experience.

Then we use all data at a particular snapshot as features, and use the price change of the equity from this day to 30 days after as label. If the price increased, it's consider a bull signal and labeled as 1. Otherwise it's considered bear signal and labeled as 0. For the case of no price change, we will treat it as bull signal. This should avoid the amount of trading.

This arrangement allows as to use supervised learning algorithm to classify the current snapshot as bull or bear signal for 30 days after. For the algorithms to choose from, we plan to use the following:

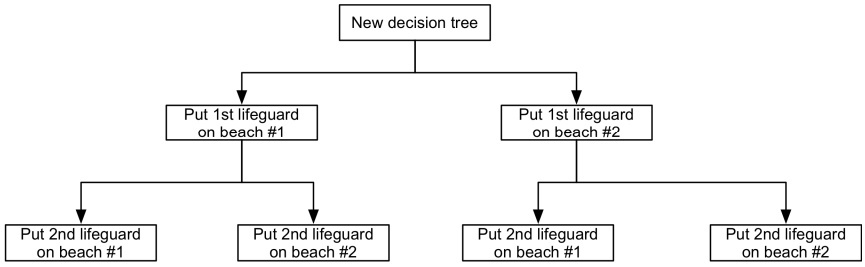
Support Vector Machines (SVM):

- SVM is good at text categorization, classification of images, hand-written recognization, and applied widely in biological science.
- SVM is effective in high dimensional spaces.
- SVM performed well in multi-dimensional dataset, which is the case here.
- SVM build a hyperplane to separate the data points into two classifications as bull and bear signal in our case. Our data set is multi-dimensional which is not as easy to visualize. For two dimensional features, SVMs work like the image below, use a line to separate data points. Image reference: wiki



Decision Tree:

- The model is easy to understand and interpret, as it uses a white box model and can see how the decisions are made.
- The decision process can be easily visualized like a flow chart.
- When the numbers of features is large, decision tree could create a model that is complex and not easy to understand.
- Decision Tree gives us analysis of important features and their weights.
- A flowchart that explains Decision Tree's decision process is as below. Image reference: wiki



Naive Bayes:

- Naive Bayes is used when the features are considered to have strong independence.
- Naive Bayes is quite to calculate and do multi-dimensional dataset well.
- If the assumption of Naive Bayes - strong independence between features, does not hold, the result could be less accurate.
- In this example, some features could affect each other. For example, low interest rate implies lower capital cost and could implies normally higher economic growth and the value of economic indicators. However, some features appear to be less directly relate to each other. For example, investor sentiments could swing to either direction dramatically and quickly, while the fundamental like economic and interest rate stay relatively the same.

For the spliting of training and test data, it makes logical sense to use past to predict future. The reverse won't be plausible anyway because future is unknown. We collected the data since year 2006, and will use duration between 2006 to 2016 as training data, and in year 2017 and 2018 as testing data.

Benchmark

Our trading suggestion will assumed to be for when to buy and sell ordinary equity such as index fund, not involve any shorting strategy. As a result, the target audience should care more about whether a predicted bull signal is actually a bull. An unpredicted bull signal, a loss opportunity, should not create too much remorse of customer. While a incorrect predicted bull signal that turn out to be bear will anger customer as it cause them to lose money. As a result, accuracy score is the better benchmark, compared to other such precision score, recall score, and F-1 score. We will use accuracy score as the benchmark of the modeling.

To compare how the strategy performs, we can compare again the passive index, that is, to use buy and hold strategy, to always buy the stock and never sell. This is equivalent to prediction that every day is a bull signal disregard of any data.

III. Methodology

Data Preprocessing

All data are read into data frame with different frequency. We plan to copy them back to the main data frame, which is the equity price data frame. We will rename the column to make them more readable when grouped together. In summary, we will perform the following cleaning:

- "data" Price data data frame: change column name from "Close" to "Price". In addition, we create new columns "PricePctP30" and "PricePctN30" to represent the percentage of price change of past 30 days change and next 30 days. We also create a label of "BULL" that based on price change of the next 30 days, to indicate the present day should be considered a bull signal or not.
- "data_tnx" 10 Year Treasury Internet Rate: copy and change column name from "Close" to "TNX"
- "data_tb3ms" 10 Year Treasury Internet Rate: copy and keep the column name of "TB3MS". The monthly data are carried forward daily until the next month's release.
- "data_lei" Leading Economic Indicator: change "ReleaseDate" to "Date" and set as index, copy and change "DeltaPrevMonthSMAShort" to "LEIMonthlyChangeSMA6". This is six month Simple Moving Average of the monthly change of the Leading Economic Indicator index. The monthly data are carried forward daily until the next month's release.
- "data_pcratio" Put/Call Ratio: copy and rename from "P/C Ratio" to "PCRatio"
- "data_aaii" American Association of Individual Investors Sentiments: copy and rename from "BullMinusBear" to "AAII". The weekly data are carried forward daily until the next week's release.
- "data_naaim" National Association of Active Investment Managers Sentiment: copy and rename from "Mean/Average" to "NAAIM". The weekly data are carried forward daily until the next week's release.
- "data_finra" Margin Debt: copy and rename from "MarginOfMarketCap" to "MarginDebtPct". The monthly data are carried forward daily until the next month's release.

When merging all the data into a single data frame, we also clip the time range of the data so all of them will have the same time duration as the "Price" data set.

Data Cleaning for avoiding look ahead bias

Given that we use the 30 days look ahead price as the data, and treat each day a snapshot, the look ahead bias is avoided, for the data of each row. Another possible look ahead bias that has to be avoided, is the training data should always be the days before test data. If we use future data to train, then look ahead bias occurs. For this, we plan to use data from 2017 and after as test data, and 2016 and before as training data. Also, we have to trim the first 30 days and last days of data as these portion of data miss information of past 30 days and next 30 days price change.

Examination of data and final cleaning

The boundary of the date range of the data has been set. We can now check if there are further problem like missing value.

NAAIM does missed value up to 2006-07-03. Given that we don't have the data, and fabricated data like using average value of it in all range does not necessary make sense. We can either drop NAAIM entirely, or avoid using those dates with missing values. Given that the missing portion about half a year, relatively small compared to the complete set of data, we decide to drop those range of dates only.

TNX, 10 Year Treasury Rate are missing on two days. These are not holiday. Another data source that provide this data, marketwatch.com are missing on these two days as well. Interest rate does not have huge daily change like equity price. Given that there are only two days, it is reasonable to use mean of the day before and after as approximate.

Benchmark Score

With our test set read, we can now calculate the score of benchmark strategy, buy and hold, e.g., assume every day is a bull signal disregard of the data.

```
Benchmark Accurarcy Score is 0.7192224622030238
Benchmark Precision Score is 0.7192224622030238
```

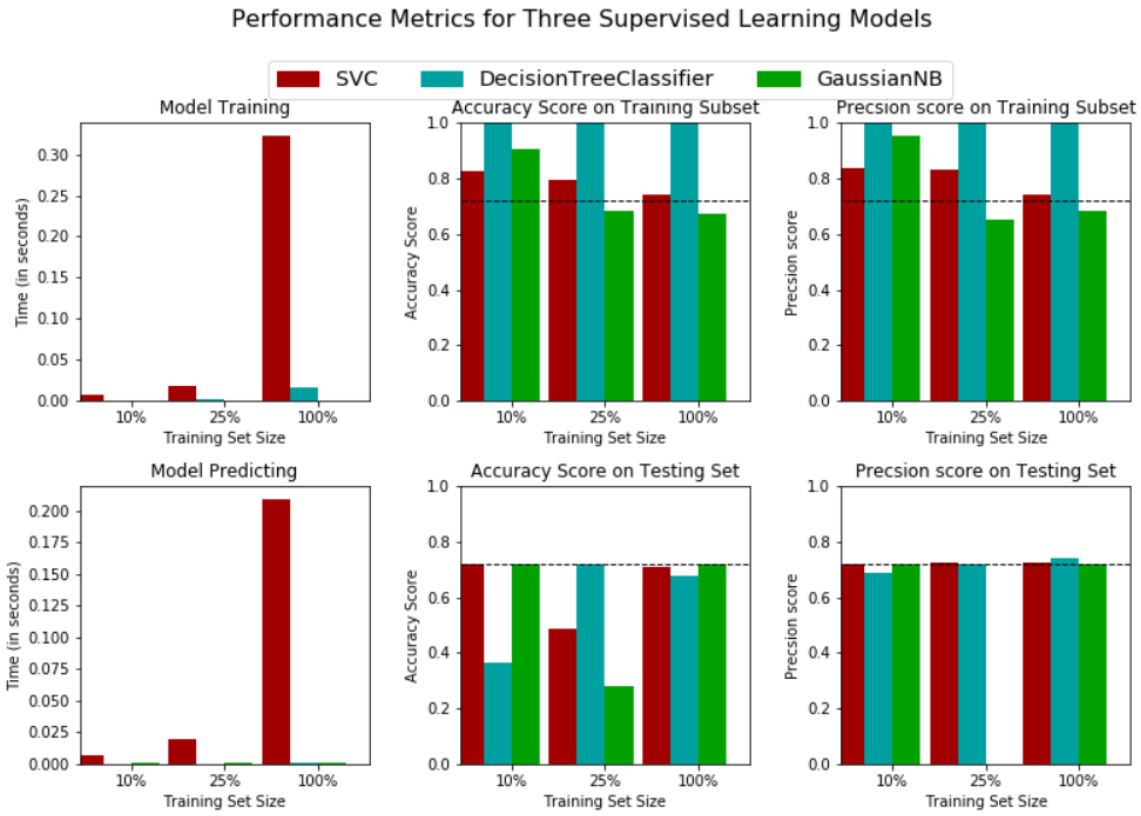
Implementation

The problem is a supervised learning problem. Among the options available, we decided on three of them to try and then refine. The algorithms chosen are:

- Decision Tree: Decision Tree execute fast and can extract the weight of the important features. In our data set, whether each feature has significant or not and how much, are from historical rule of thumb. There is no truth based evidence whether they actually has significant. Decision Tree process will help identify how important each feature is.
- Support Vector Machine: SVM's hyperplane approach deal with multiple dimensions well. It's flexible nature should make the fitting better and easier to fine tune into not over and under fit.
- Naive Bayes: Again, our data set does not have assumption about dependency between them. Although they could be dependent, they could still be sufficiently independent that fits to Naive Bayes' assumption. It also has the benefit of speedy execution.

For the scoring metric. Accuracy score is the best as it score whether we identify both bull and bear signals correctly. If we can identify both bull and bear signals accurately, we can use the data to identify for both long/short trading strategy. Another possible score is precision score. In a more limited use case, where we only provide bull signal, and user uses that for long trading strategy only, not as a short strategy, then a higher precision score would be more beneficial for such trading strategy.

Given this we will use accuracy score as major metric, while using precision score as additional reference.



Refinement

Training set are further sampled into either 1%, 10%, and 100% of training data respectively to see how the models perform with reduced size of training set. Interesting we found the SVM method will throw error that there is only one class in "y". Although this is allowed in Decision Tree and Naive Bayes algorithms, it is indeed not reasonable when the sample data set are either all bull and bear signals. By choosing the reduced sample sizes to 10%, 25%, and then full 100% respectively, the abnormality of training set sample are avoided.
All three models perform well in testing, with when 100% of the training data are used. When training size are only 10% or 25% of total training data, all three models show weakness with a lot lower scores, especially Decision Tree and Naive Bayes. SVM's lower score show up in 25% sample size, but not as bad as the other two models.

Decision Tree appears to have some overfitting problem with it's extremely high training score. SVM and Naive Bayes have less overfitting problem and still give reasonable test score on both accuracy and precision.

In terms of the most important benchmark of accuracy score, and important in long only trading strategy of precision score, all three models are on par without huge deviation between them.

Given above evaluation, overall, SVM appears to be good recommendation given it's more robustness if different sample size, and stable performance of benchmark accuracy and precision

score, without suffering from big overfitting problem, it should be the recommended strategy. And we choose it to optimize.

Although we do not intend to recommend Decision Tree, it does perform well despite overfitting. And give us insight into important features. This definitely will have a lot of interests. One reason is that we used words of wisdom and choose various macro economic data and sentiments data as features, whether any of them have significant is unknown, or in other word, not machine proved. It will be interesting just to see what important features are as seen by Decision Tree algorithm. Hence we also decide to optimize it.

For SVM model:

```
Unoptimized model
------
Accuracy score on testing data: 0.7192
Precision on testing data: 0.7192

Optimized Model
------
Final accuracy score on the testing data: 0.6544
Final precision on the testing data: 0.6998
```

We can observe that the final tuned model use the paramters of:

- kernel: poly
- C: 1.0
- decision_function_shape: ovo
- degree: 2

The fine tuning using grid search appear not able to make the score any better. Both accurarcy and precision score stay at 0.7192

For Decision Tree Model:

```
Unoptimized model
------
Accuracy score on testing data: 0.6760
Precision on testing data: 0.7414

Optimized Model
------
Final accuracy score on the testing data: 0.7192
Final precision on the testing data: 0.7192
```

IV. Results

We narrowed down to two algorithms SVC, and Decision Tree. With SVC appears better in its ability to generate stable result even with small sample size. Decision Tree can give us some insights into important features. Both models are tuned to attempt to improve.

For SVC, we used grid search method to try on the following hyper parameters:

- kernel: linear, poly, rbf
- C: 1.0, 0.1, 0.5, 10.0, 100.0
- decision_function_shape: ovo, ovr
- degree (for poly): 2,3,4,5

Interestingly, none of these are able to improve SVM any further, with accuracy score and precision score actually both reduced slightly from 0.7192 and 0.7192, to 0.6544 and 0.6998

For Decision Tree, we also used grid search for the following hyper parameters:

- max_depth: 1,2,3,4,5,None
- max_features: 1,2,3,4,5,6,7
- min_samples_split: 2,3,4,5
- min_samples_leaf: 1,2,3,4,5

Were able to improve accuracy score from 0.6760 to 0.7192, while the precision score degraded a bit from 0.7414 to 0.7192.

Recall that the benchmark, the famous buy and hold strategy of passive investment has both accuracy and precision score of 0.7192, we found that we can only match, but not able to beat the benchmark.

Model Evaluation and Validation

We perform a few validations of the model by further splitting the data and change the initial state of the model.

Validation by random state

To further evaluate whether, we can evaluate them with different random states, and review their resulting scores under different random states. Previously we have been using 44 for all models. We can further check if random state could affect the scores greatly to check if the models have a high variance.

We do this by change the previously best fit models by retaining all the parameters except for the random state, where we use five different values.

The result on Decision Tree is as follow:

	Accurarcy Score	Precision Score
44	0.719222	0.719222
123	0.719222	0.719222
9999	0.719222	0.719222
1000	0.719222	0.719222
1	0.719222	0.719222

The scores do not change over different random states, which validate that the model is stable.

We do the same for SVM, keeping the previously found best hyper parameters, and change random states only and re-train and test the models.

	Accurarcy Score	Precision Score
44	0.654428	0.654428
123	0.654428	0.654428
9999	0.654428	0.654428
1000	0.654428	0.654428
1	0.654428	0.654428

The results of SVM under different random states aslo show the model is stable as the score does not vary by the difference in random state.

Validation by time series data set split

One technique of splitting data, as we have done earlier, is by splitting the data at certain date. And using the data before that date for training, and after that date for testing. This article [8] has further details. We can reuse this technique and shrink the training set to be smaller, so we have enough data to split between them. For our data set, previous we use data from 2006 to 2016, to test the year of 2017 and 2018. We can reduce the training set and use 2006 to 2014, to test 2015-2016, use 2006 to 2012 to test 2013-2014, and so on.

For Decision Tree Model:

	Accurarcy Score	Precision Score
Train 2006 to 2008, Test 2009 to 2010	0.660714	0.660714
Train 2006 to 2010, Test 2011 to 2012	0.657371	0.657371
Train 2006 to 2012, Test 2013 to 2014	0.769841	0.769841
Train 2006 to 2014, Test 2015 to 2016	0.646825	0.646825
Train 2006 to 2016, Test 2017 to 2018	0.719222	0.719222

The scores using these data split on the Decision Tree model show the models are relatively stable, but have lower scores during the bear market or stock correction years of 2015-2016, where Fed reserve stopping quantitative easing and China economic slowdown caused some overall worse performance of stock market. 2011 has Europe dual recessions also is one of the significant during the whole time period of training and testing.

For SVM model:

	Accurarcy Score	Precision Score
Train 2006 to 2008, Test 2009 to 2010	0.488095	0.488095
Train 2006 to 2010, Test 2011 to 2012	0.647410	0.647410
Train 2006 to 2012, Test 2013 to 2014	0.636905	0.636905
Train 2006 to 2014, Test 2015 to 2016	0.557540	0.557540
Train 2006 to 2016, Test 2017 to 2018	0.654428	0.654428

The scores using these data split on the SVM model show the models are relatively stable, but have lower scores during the bear market or stock correction years of 2015-2016, and 2011. The lower score is more significant than Decision Tree.

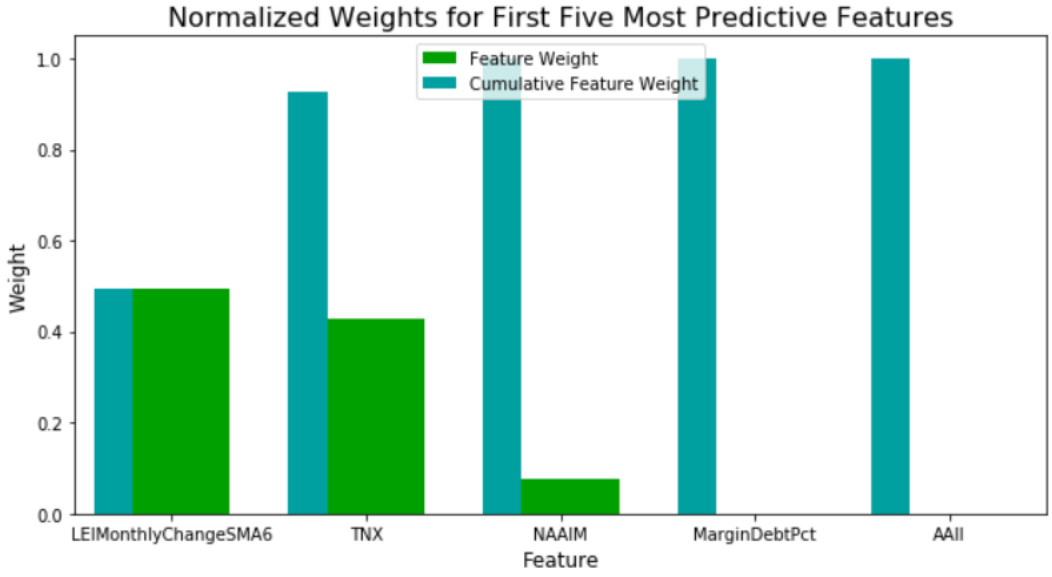
Justification

Both Decision Tree and SVM perform well and stable under different random state. When using different sample size by selecting different period of back test period, the higher scores are achieved by Decision Tree compared to SVM in all cases. This suggest Decision Tree is better algorithm in this particular problem. The Time splitting technique does reveal the well known issue in this domain, or maybe for all time series data. That is, when using certain amount of history as training data, which period of history you look at make a difference. Sometime a huge different. Also, the market is known to have known cycles[4], which could have frequency of a few years. That translates to stock market having cycles between bull and bear market, which bull market average 9.1 years and bear market average 1.4 years[9]. When the test period coverage the bear market, the past strategy tends not to work. And it does reveal in lower scores during bear years. However, all scores still achieve higher than random score (0.5) and provide reasonable guideline for investors.

V. Conclusion

Free-Form Visualization

Previous we use Decision Tree as one of the model to investigate and improve, it also provides us with the features that are considered important.



Reflection

On the result

Our study found that we can match, but not able to beat the benchmark strategy of buy and hold, passive investment. Although it is somehow disappointing, it is actually quite remarkable findings as it shows the difficulty of reliably beating the market might actually be a mirage and the strategy that actually beats the market, but be a form of overfitting that is bound to cause problem in the real world application. In a way, it proves that the study and philosophy of the random walk strategy[5], the type of strategy the usually comes from academic rather than industry participant, does have some merit.

The important features analysis, reveal that only three features are considered important out of the total of seven we tested. The in order of most significant, are:

- Six month moving average of the change of Leading Economic Index from previous month. (49%)
- 10 Year Treasure Interest Rate (42%)
- Investment Sentiment Survey of National Association of Active Investment Managers (7%)

The finding does make very intuitive sense in that the old saying that market is not the economic, but are indeed resemble the economic. Also, bond market has always been the most direct competition of stock market for investor's money, and given 10 Year Treasury is usually considered 'risk free' return, and being the floor of how financial assets are priced, it's interest rate definitely plan a big role in determining the market value.

Sentiment is a form of investor psychology and does move the market in some way. Although not a big factor.

On the process

The data collecting process is challenging as they are coming from various of organizations and have their own different format, different frequency. There are some for profit providers of financial data to provide services in this area. The fact that these for fee service exist explain the challenging of collecting the data.

One we have the data ready, the one important constraint of time series data is to avoid look ahead bias, that is, you cannot use future data to train and predict past data, as this violates real world logic. As a result, techniques of splitting training and testing data cannot be arbitrarily applied. We are forced to split the training and testing data with strict separation of a certain cut off date.

In this research, we use a novel approach of using a snapshot features in a particular day to predict future days. This gives us a large amount of data to train while still able to avoid look ahead bias described in previous paragraph. However, this approach also limits us in use features for one day only, instead of a certain period of past history, for example, past 30 days.

Improvement

- Seven features were collected and tested. Although they might be representative enough, there are way more available and could be more significant data that can be gathered and studied.
- Instead of looking at one day snapshot, we could use a period of time and then use aggregate function like mean as features input. There could be decision on where to use rolling window for this period that will provide larger sample size, or use non-overlapping window which could reduce sample size the larger the period is. The period approach makes some logical sense as the dynamics in the history could make a different. Although one could argue snapshot is best as the features for that day have been 'priced-in'.
- We also attempted to generate only a boolean signal of bull or bear. We could expand that signal into range of return, for example, top quartile of positive to bottom quartile of negative return. This provides more classes and could also add investor in having an expected return value to help them decide on investment strategy.

VI. References

- [1] Fama, E. F.; French, K. R. (1993). "Common risk factors in the returns on stocks and bonds". Journal of Financial Economics.
- [2] Fama, E. F.; French, K. R. (1992). "The Cross-Section of Expected Stock Returns". The Journal of Finance.
- [3] Shiller, Robert (2005). Irrational Exuberance (2d ed.). Princeton University Press. ISBN 0-691-12335-7.
- [4] Dalio, Ray. A Template For Understanding Big Debt Crises. (2018)
- [5] Malkiel, Burton Gordon (1973). A Random Walk Down Wall Street: The Time-tested Strategy for Successful Investing. New York: W.W. Norton. ISBN 0-393-05500-0
- [6] André Kostolany (1996). Weisheit eines Spekulanten - German
- [7] Renuka Joshi (2016). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures
- [8] Francesco (2014). Pythonic Cross Validation on Time Series
- [9] First Trust (2018). History of U.S. Bear & Bull Markets Since 1926 (https://www.ftportfolios.com/Common/ContentFileLoader.aspx?ContentGUID=4ecfa978-d0bb-4924-92c8-628ff9bfe12d)