**CS 6075.501 Machine Learning**

**Final Report**

**Introduction**

The project attempts to capitalize on stock selection strategies generated by machine learning learners and then test the strategy. The use of mathematical models to replace the traditional quantitative trading main point of view, based on the strategy of computer technology, but it is still rare, the application of machine learning in the field of quantitative transactions. In hence, this project is also an exploration in this area. Now, the machine learning process, there will always be good results, the test of future investment, there is no such a satisfactory result.

Here we assume that we have a fund worth 1 million, and then we have annual data of 50 stocks from 20170104 - 20171201, the first 80% of the data will be used to train learners. Then learning will be based on today's data given the second 20% of tomorrow's stock price forecast. We rely on the forecast to do the transaction. Finally, the yield is compared with the transactions based on the S&P 500 index to see if the strategy will result in more profits.

**Related Work**

At present, machine learning has been widely used in the fields of image recognition and automatic driving. However, in the financial field, machine learning is seldom used in quantitative trading, and we can say that this result is far from satisfying. In the data is relatively fixed, enough data, with high signal-noise ratio, machine learning has good results. But for the quantitative trading, the opposite is true, at the extreme, the financial community has no rules. We still tend to describe financial data linearly and qualitatively. Machine learning can give over-fitting results. In the lead Récek thesis, they used a mixture of Gamma and Gaussian distribution models to determine the optimal trading strategy [1]. Since they are based on a semiparametric distribution, the bottleneck still exists in the result, and the key lies in the choice of distribution.

At the same time, some avantgarde financing institutions, such as Deutsche Bank, have already begun to use machine learning based models named N-LASR (Nonlinear Adaptive Style Rotation) models [2]. In this project, our primary focus is on the accuracy analysis of strategies obtained from multiple learners through backtesting. Although we may get good results in the backtesting, there is not so satisfactory result in the future practical application. While data processing capabilities are increasing day by day, the combination of machine learning and traditional quantitative trading methods will surely bring us hope.

**Dataset Description**

This project is based on the dataset provided by Yahoo finance. The dataset provides series of data for the 500 typical stocks that contained in the S&P 500 indexes in United States. We choose the highest value, close value, lowest value and trading volume of these 50 stocks from 01/04/2017 - 12/01/2017 as the basic data pool.

Thanks to the Python package TA-Lib [3] that have done most of the calculation, we get 8 features as parameters for the learners to get our strategies. The features are as followed:

1. Exponential moving average (EMA) [4] EMA are commonly used in conjunction with other indicators to confirm significant market moves and to gauge their validity(using the function ta.EMA).

2. Moving average (MA) [5] is a widely used indicator in [technical analysis](https://www.investopedia.com/terms/t/technicalanalysis.asp) that helps smooth out [price action](https://www.investopedia.com/terms/p/price-action.asp) by filtering out the “noise” from random price fluctuations, being used to identify the trend direction and to determine support and resistance levels (using the function ta.MA).

3. Momentum [6] (MOM) is to compare the current price with the previous price from a selected number of periods ago (using the function ta.MOM).

4. Rate of change(ROC) [7] is a method to measures the percent change in price from one period to the next.  Calculated by the formula ((price/prevPrice)-1) \*100. (using the function ta.ROC)

5.Weighted Close price (WCLPRICE)[8] is one of [technical analysis](http://www.ta-guru.com/Book/TechnicalAnalysis/TechnicalAnalysis.php5) indicators which averages price of each period, like [typical price](http://www.ta-guru.com/Book/TechnicalAnalysis/TechnicalIndicators/TypicalPrice.php5) and [median price](http://www.ta-guru.com/Book/TechnicalAnalysis/TechnicalIndicators/MedianPrice.php5) (using the function ta.WCLPRICE)

Weighted close is calculated by formula:

ormula for Weighted Close

6. Normalized Average True Range (NATR) [9] attempts to normalize the average true range values across instruments(using the function ta.NATR) by using the formula( ATR(n) / Close \* 100)

7. Standard deviation (STDDEV)[10] is a statistical term that measures the amount of variability or dispersion around an average also a measure of volatility (using the function ta.STDDEV)

8. The Weighted Moving Average (WMA) [11] places more emphasis on recent prices than on older prices. Each period’s data is multiplied by a weight, with the weighting determined by the number of periods selected (using the function ta.WMA).

**Methods**

**Coding language**

Python (2.7.14)

**Data preprocessing methods**

First, we choose 50 stocks from S&P500, saving their stock code as a text file. We read every stock code line by line and download its data from Yahoo finance. Then we choose 8 features from ta-lib which provides many stock performance indices and process to create data of these 8 features for these 50 stocks. For all stocks, with the close price of original data, we need to calculate the price change rate of each day by comparing to the last day and give labels,1 and 0, to each day judged by if the price change of that day is larger than 0 or not. Then we save all data tables of these 50 stocks to a list, the first 80% as training data, the left 20% as testing data.

**Processing methods**

Use SVM, ANN, Naïve Bayes, LR and KNN 5 methods to training all data of 50 stocks. Then use every method to predict the labels of testing data, the data of left days of 50 stocks (20%). If the label is 1, we regard the price of that stock will arise on that day. And also, we will count how many stocks which label is 1 on that day, dividing original fund 1000,000 to those stocks and calculate the benefits of that day. Finally, we obtain all benefits of the test days of 5 methods.

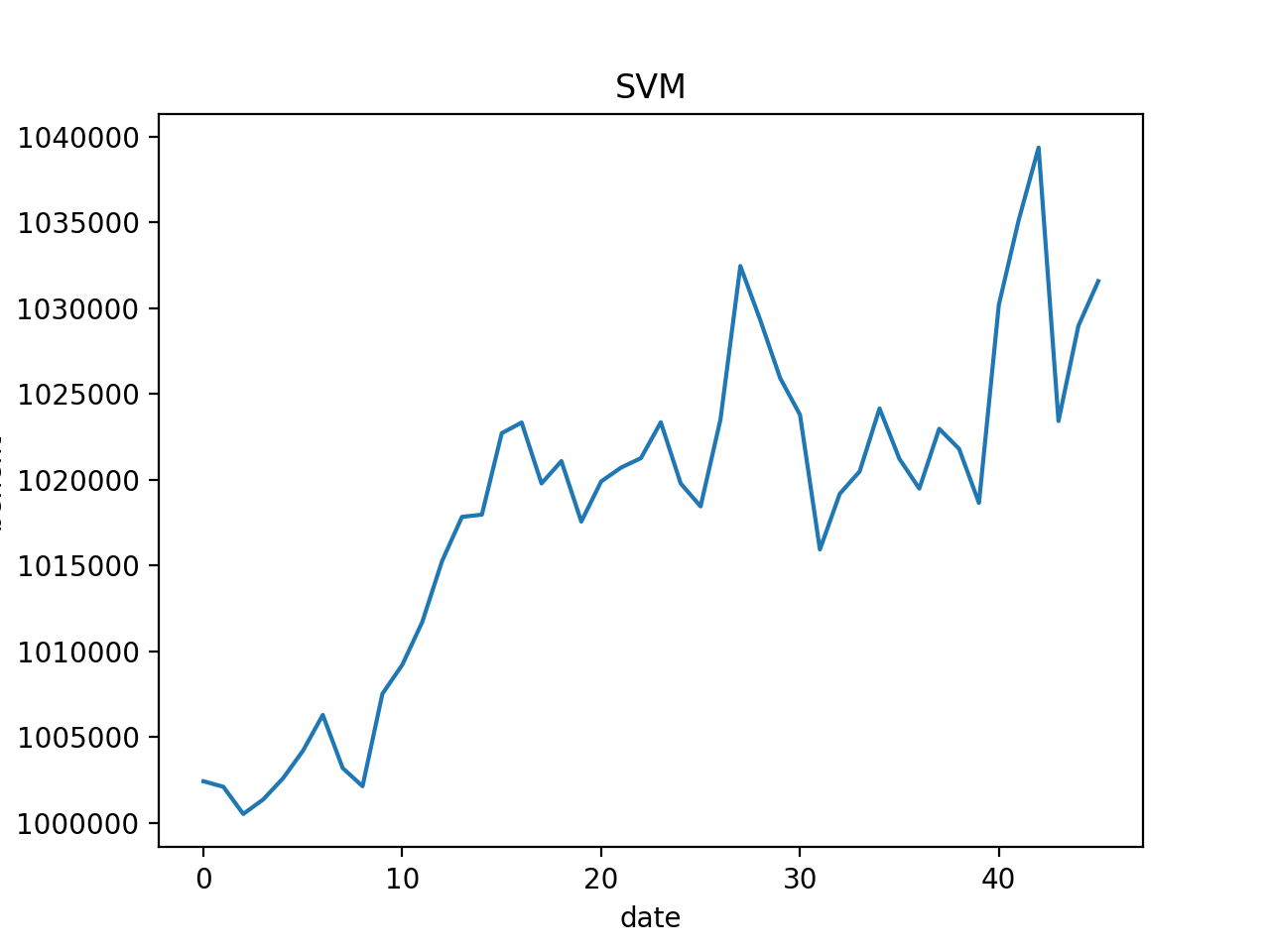
The important thing is how to evaluate the results of 5 methods. We need to calculate benefits of testing days through S&P500. Then we can see if there are some learning tools are better than the market standard analysis index.

**Experimental Evaluation**

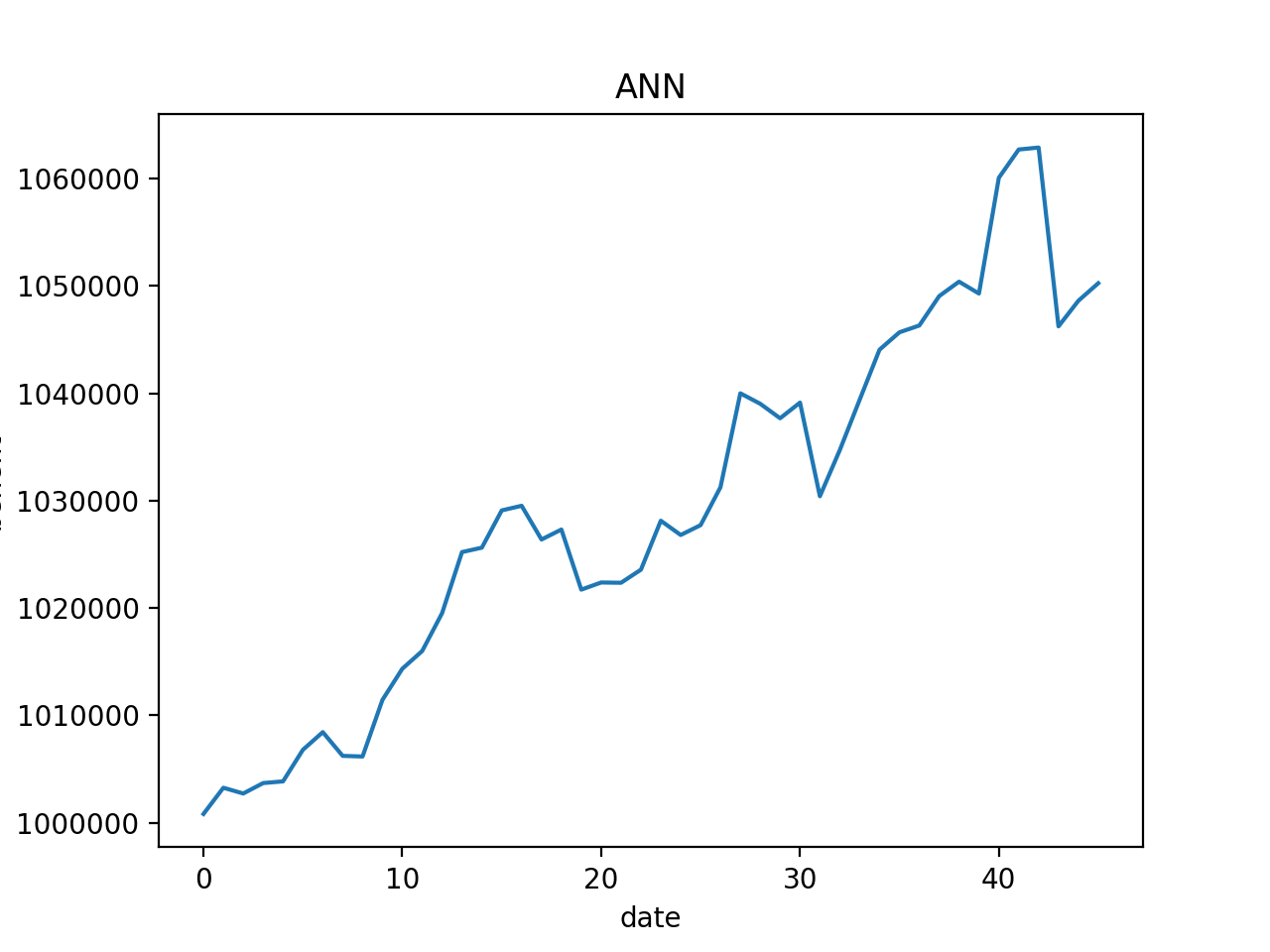
**Results**

We have used five classifiers to try those data. The five classifiers are SVM, ANN, Naïve Bayes, LR and KNN. Result for the five classifiers are as followed:

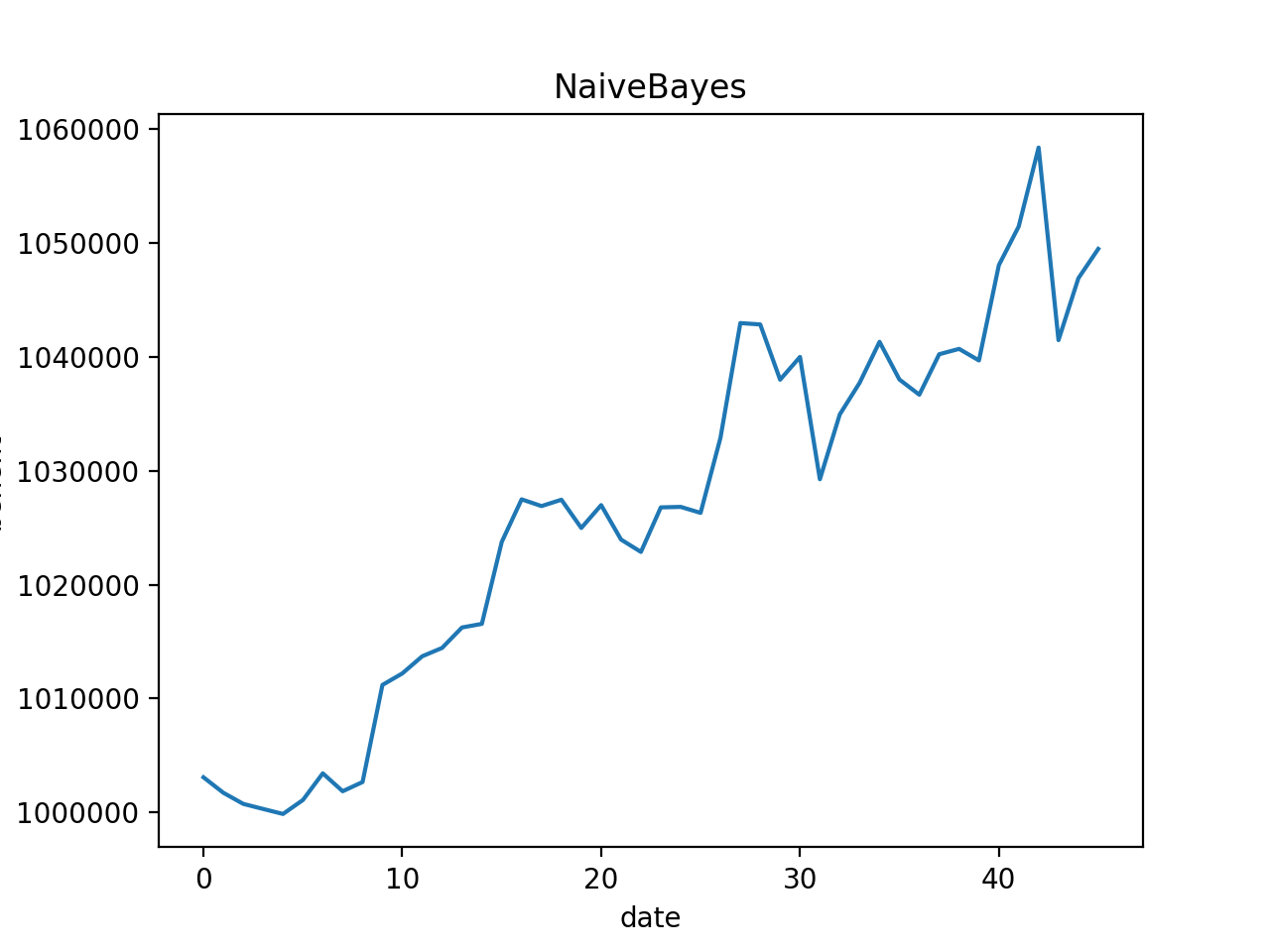
1. SVM



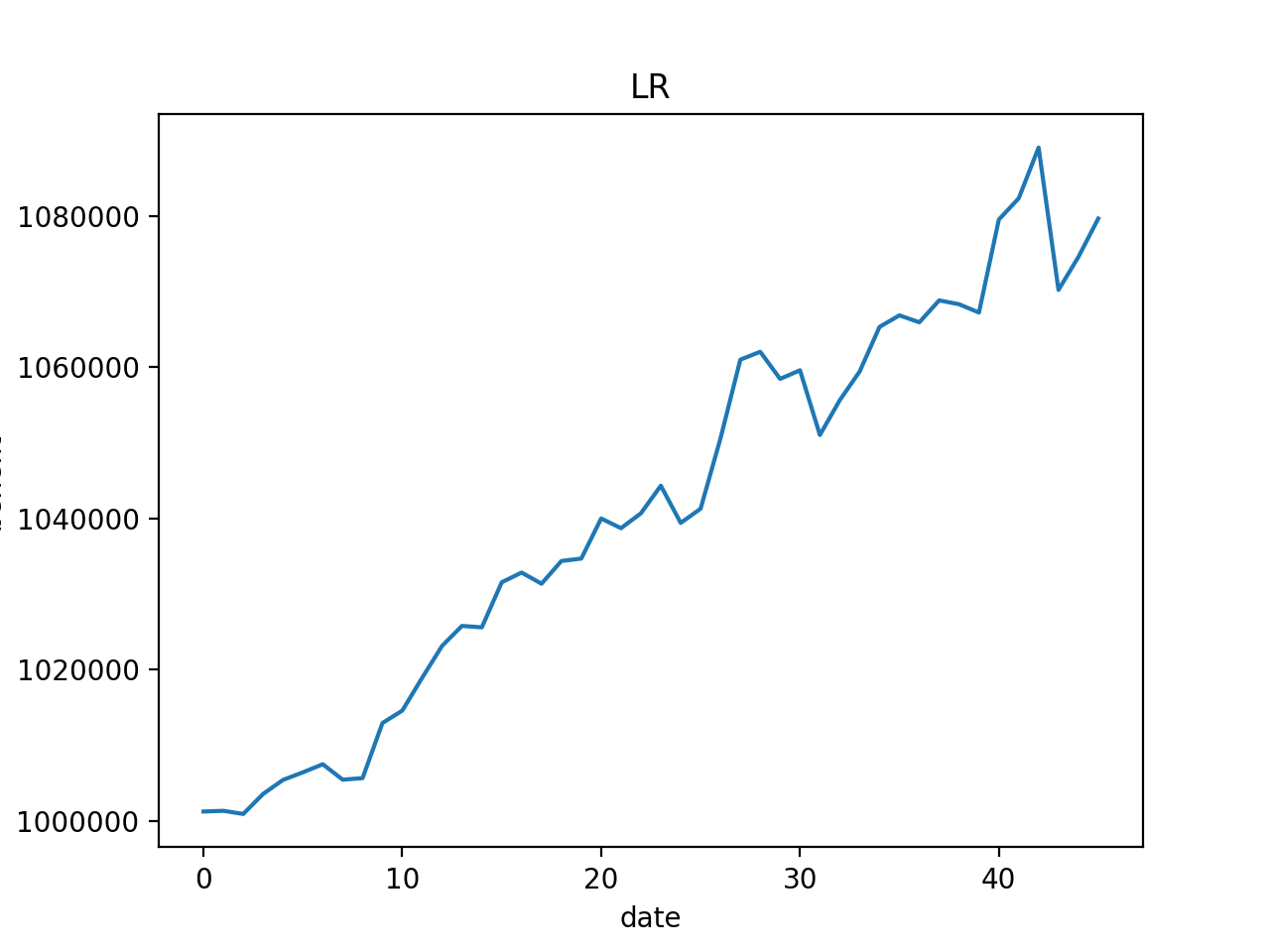
1. ANN



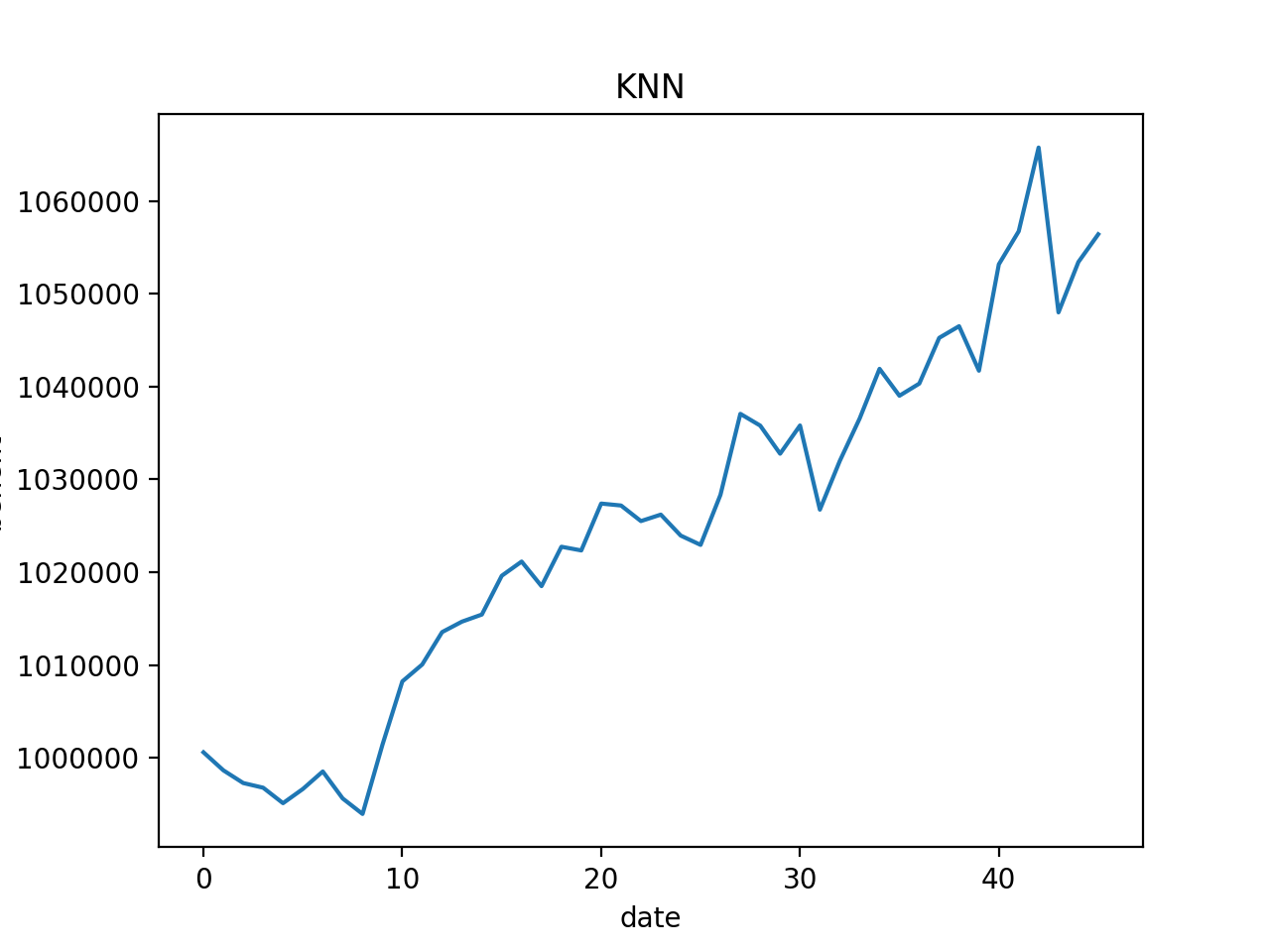
1. Naïve Bayes



1. LR

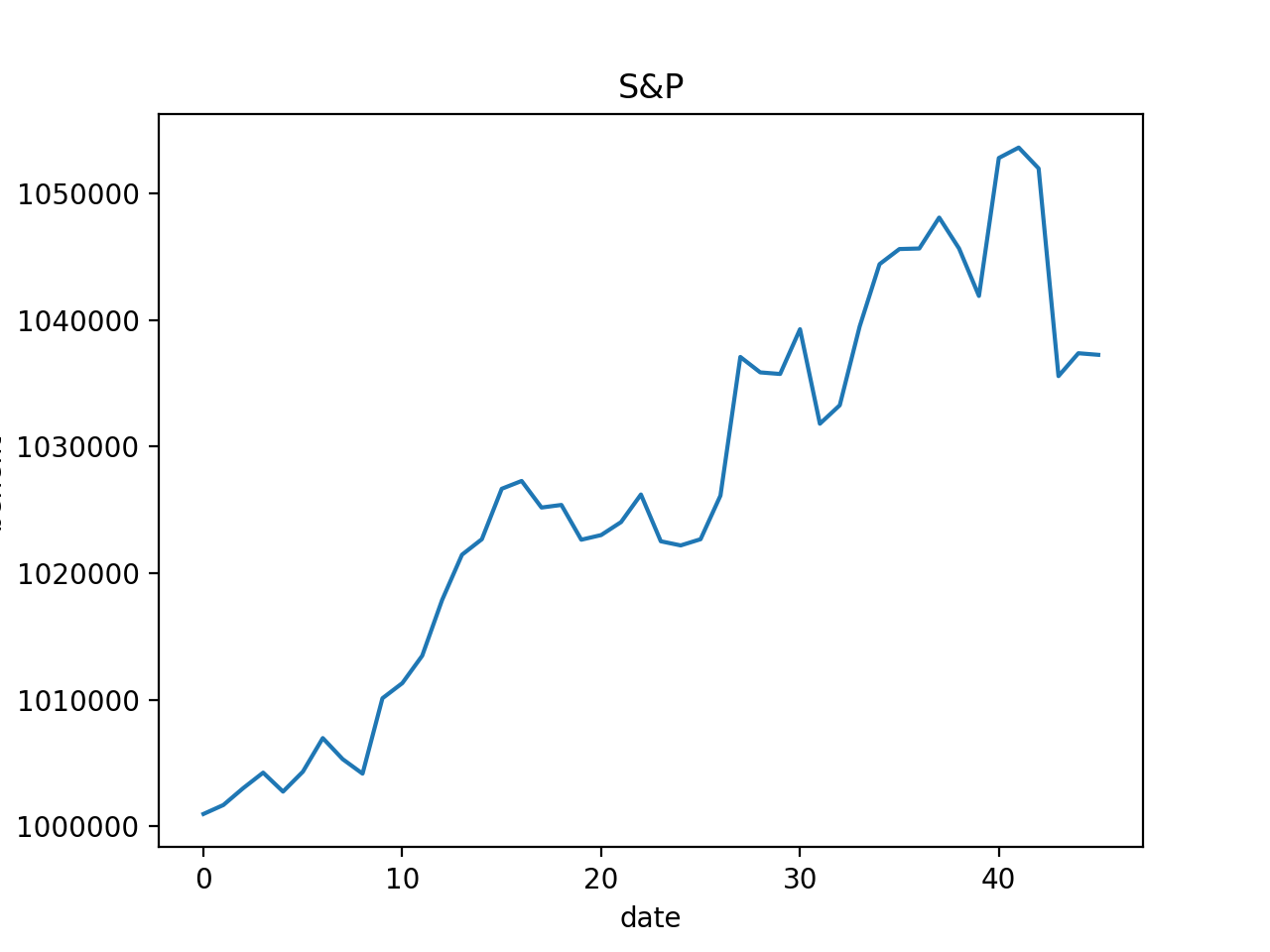


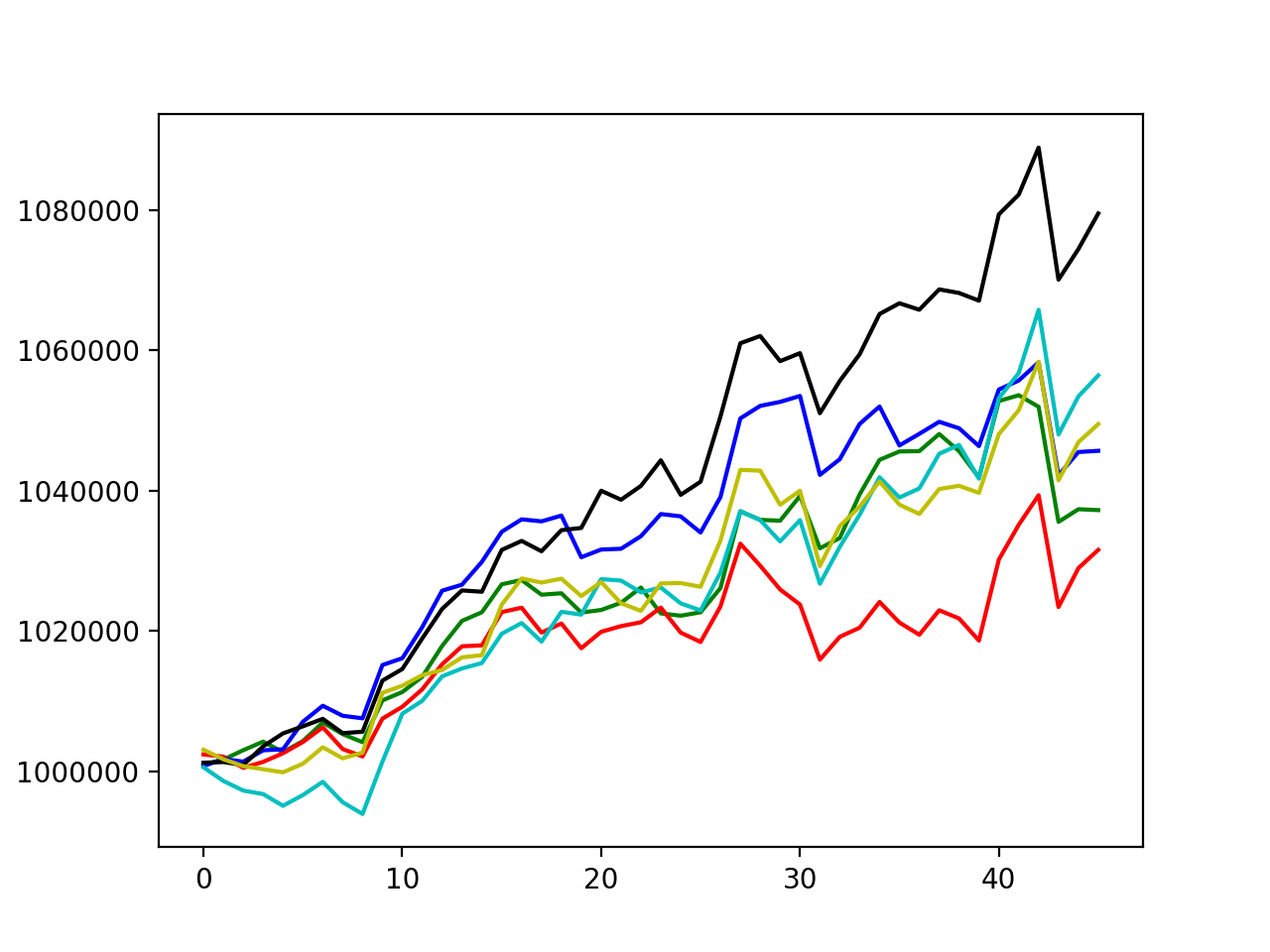
1. KNN



To evaluate the result of these 5 classifiers, we also give the benefits of S&P 500 of same days. Also, we put those results together to compare the benefits change of each day.

1. S&P 500



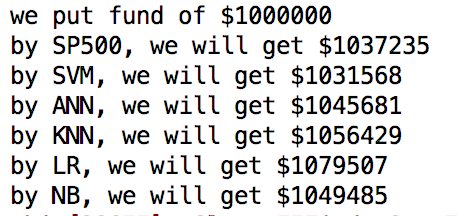


Green: S&P 500

Red: SVM Blue: ANN Yellow: Naïve Bayes Marine: KNN Black: LR

**Analysis**

From all the graphics showed above we can get that four of our hypothesis are supported. The separate analysis for each of the five classifiers are as followed:



We invest 1000,000 dollars for last 20% days of this year, and through 5 machine learning tools, LR will help us to earn the most money, **42,272** dollars more than standard S&P500

**Conclusion**

This project proves that for the five classifiers we used, the LR gives the best result so we take it as a good training tools for stock market. But these are only 5 training tools in thousands training tools. We need to analyze and give experiment to more classifiers. We can also see SVM is worse than the standard S&P 500. And this result may be caused by the data bias of the 50 stocks we chose. choosing which stock to be predicted is a significant condition. A better method to avoid this situation is to choose as more stocks as possible for training data. Another way to improve the result is to choose more representative features as parameters. Not only the past data can affect the future price of stocks in market, but also the news and public sentiment will also have an influence.

**References**

1. Lead Rezek, “Constrained Mixture Models for Asset Returns Modelling”, 03/15/2011

2. Sheng Wang, Miguel-A Alvarez, Javed Jussa, Zongye Chen, Allen Wang, Yin Luo, “The wisdom of crowds: crowdsourcing earnings estimates”, Deutsche Bank, Markets Research, 03/04/2014

3. TA-Lib: Technical Analysis Library, http://ta-lib.org/index.html, Web, 04/30/2017

4.EMA: [Exponential Moving Average (EMA)](https://www.investopedia.com/terms/e/ema.asp#ixzz50Exn8WTC) <https://www.investopedia.com/terms/e/ema.asp#ixzz50Exn8WTC>

5.MA: [Moving Average (MA)](https://www.investopedia.com/terms/m/movingaverage.asp#ixzz50Ewa0M00) <https://www.investopedia.com/terms/m/movingaverage.asp#ixzz50Ewa0M00>

6. MOM: <https://www.tradingtechnologies.com/help/x-study/technical-indicator-definitions/momentum-mom/>

7. ROC: <http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:rate_of_change_roc_and_momentum>

8.wclprice

<http://www.ta-guru.com/Book/TechnicalAnalysis/TechnicalIndicators/WeightedClose.php5>

9. NATR:

<http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:standard_deviation_volatility>

10. WMA:

<https://www.tradingtechnologies.com/help/x-study/technical-indicator-definitions/weighted-moving-average-wma/>