# Forecasting S&P 500 Index Movement Direction with Machine Learning Algorithms

Xin Jing
xin.jing@mail.utoronto.ca

May 7, 2017

**Abstract** Machine learning algorithms has been used to find patterns in historical financial data, in anticipation of predicting future changes of the stock market. In this project, I examine two machine learning methods utilized to predict the future direction of S&P 500 Index movement. Specifically, these two methods are Support Vector Machine(SVM) and Logistic Regression.

## 1 Introduction

We have seen in previous literature the usage of various financial time series as features to predict future movement of the stock market. These features usually include the past values of the stock of interest, the trading volumes of these stocks, the stock indices from other parts of the world, and other macroeconomic variables such as GDP, Volatility Index, Uncertainty Index, performance of the currency used in that country, etc.

Despite the wide rage of features used in previous experiments, few of the successful experiments excluded the past values of the stock itself as a predictor. This suggests that the past values of the stock may possess significant predictive power of the future values, and it casts doubt on the predictability of other predictors, since if other predictors do have influence on the stock price, the past values of the stock price should carry enough information from other predictors. Therefore, I want to examine if we can use the lagged values of the stock itself to achieve the same level of predicting accuracy that was achieved in previous experiments.

In Part I of this project, I will predict S&P 500 Index movement with the percentage changes of its past values. In Part II, out of personal curiosity, I will examine the predictive power of the lagged percentage changes of S&P 500 Index, together with the lagged percentage changes of Volatility Index (VIX). This is because the two variables of the same time period are highly negatively correlated, with correlation coefficient -0.7. In addition, according to Sun, 2008[1], adding volatility index will increase the predictability of stock market returns.

## 2 Experiment Design

I use historical weekly data of S&P 500 Index and VIX from Yahoo! Finance. The range of the data is from 2000/4/26 to 2017/4/24. Therefore, there are 887 observations. In both parts of the project, the response variable is the direction of S&P 500, with 1 and -1 denoting upward and downward movement. In Part I, the predictors are the lagged change rates of S&P 500 Index. The model can be written as the following function:

$$Direction_t = F_1(C_{t-1}^{S\&P500}, C_{t-2}^{S\&P500}, C_{t-3}^{S\&P500}, ..., C_{t-p}^{S\&P500}), \tag{1}$$

where $Direction_t$ takes either 1 or -1, and $C_{t-i}^{S\&P500}$ denotes the $i_{th}$ lagged percentage change of S&P 500 Index. I will experiment with both Support Vector Machine and Logistic Regression.

In Part II, the predictors include the past change rates of S & P 500 Index and those of VIX. The function can be written as:

$$Direction_t = F_2(C_{t-1}^{S\&P500}, C_{t-1}^{VIX}, C_{t-2}^{S\&P500}, C_{t-2}^{VIX}, C_{t-3}^{S\&P500}, C_{t-3}^{VIX}, ..., C_{t-p}^{S\&P500}, C_{t-p}^{VIX}), \tag{2}$$

where $C_{t-i}^{VIX}$ denotes the $i_{th}$ lagged percentage change of VIX. I will use Support Vector Machine in this part only, since as will be shown in the first part, Support Vector Machine outperforms Logistic Regression in this classification problem on the nonlinear data set.

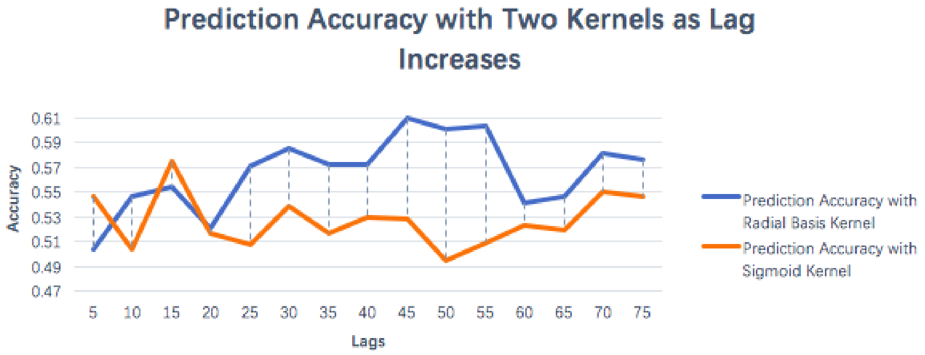
In both parts of the project, I will look at the prediction accuracy (the percentage of correct predictions) and find the best lag value with the best training size. The implementation of the algorithms will be done using scikit-learn package in Python.

# 3 Models & Results

# Part I

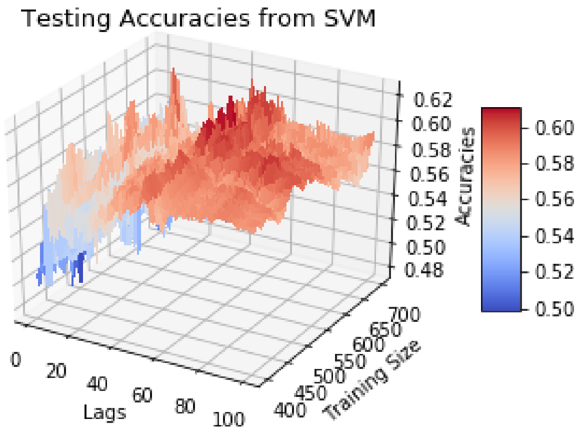
## 3.1 Support Vector Machine

**Choice of Kernels** I experimented with linear kernel, radial basis kernel(RBF), and sigmoid kernel. Due to the non-linearity of the data set, linear kernel does not work for the project. To determine which of the other two kernels is better-suited for this problem, I used an idea similar to cross-validation (cross-validation is not suitable for time series because the order of the training set and testing set matters in time series). I divided the first 800 observations into 4 subsets, each with 200 observations. The 4 subsets are used as 4 separate training sets, and the corresponding testing sets are the next 60 observations right after each of the training set. For each kernel, I examined the average prediction accuracy with different lags of past change rates. The results are shown in the following figure:



As can be seen from the chart, as lag increases, the RBF kernel outperforms the sigmoid kernel. Since larger lags with RBF kernel produce better accuracy, I choose RBF kernel.

**Find the Best Lag with the Best Training Size** After we've decided to use RBF kernel, we'll look for the combination of lag and training size that gives the highest prediction accuracy. To full utilize the data, for each training set of a given size, we use the rest of the data points as the testing set. Therefore, as lag and training size increase, the testing data will be less. To keep the testing set representative, we want to control the lag and training size. I choose lags to be from 1 to 100, and training sizes from 400 to 700. The 30,100 results are shown as 3D plot below:



No matter what the training size is, the testing accuracy peaks when the lag is around 50. The best accuracy is 62.7%, generated by the combination of lag 52 and training size 606. The following is some descriptive statistics of the accuracy produced by different lags and training sizes using SVM with RBF.

kernel:

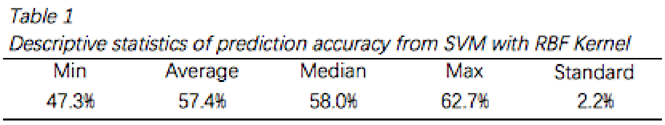


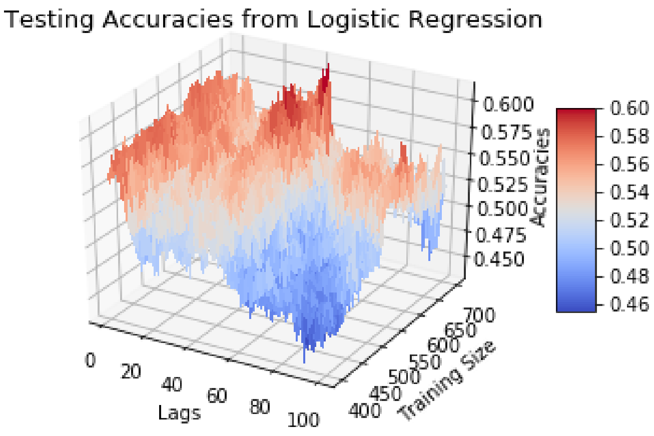
*Table 1*

*Descriptive statistics of prediction accuracy from SVM with RBF Kernel*

| Min | Average | Median | Max | Standard |
|---|---|---|---|---|
| 47.3% | 57.4% | 58.0% | 62.7% | 2.2% |

## 3.2 Logistic Regression

Same as above, we fit logistic regression with past change rates of S&P 500, and we aim to find the best lag of the change rates and the best training size. Here, we use L1 regularization in optimization. The results are shown in the picture below:



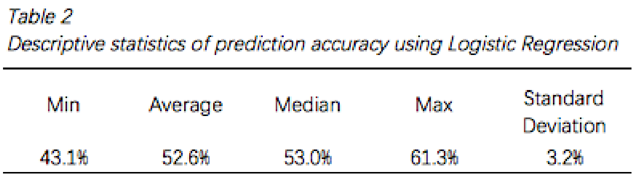


*Table 2*

*Descriptive statistics of prediction accuracy using Logistic Regression*

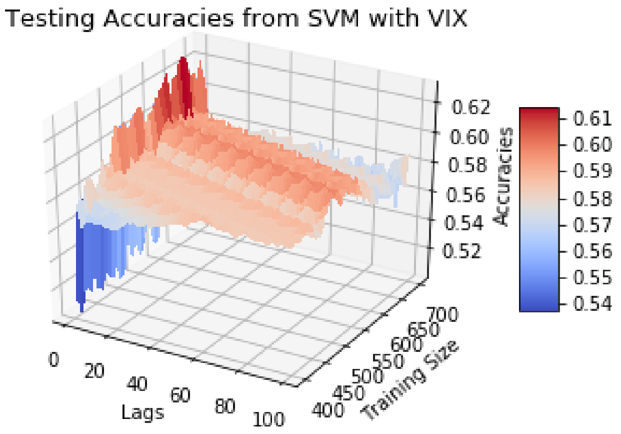
| Min | Average | Median | Max | Standard Deviation |
|---|---|---|---|---|
| 43.1% | 52.6% | 53.0% | 61.3% | 3.2% |

The best accuracy is 61.3%, given by lag 45 and training size 700. However, as can be seen from the descriptive statistics, the average accuracy is way below that from SVM method.

Logistic regression does not perform as well as SVM. Intuitively, this is due to the non-linearity of the data set. Logistic regression is a linear algorithm, which falls short of the goal of modeling non-linear data set. For SVM, thanks to the flexibility of utilizing different kernels (linear and non-linear), SVM achieves better performance on non-linear data set than logistic regression does.

## Part II

In Part II, we'll predict the future movement direction of S&P 500 using not only its lagged change rates, but also the lagged change rates of VIX, as is given by function (2). Since SVM performs better on non-linear data set, we'll use only SVM with RBF kernel in this part. The results are given below:



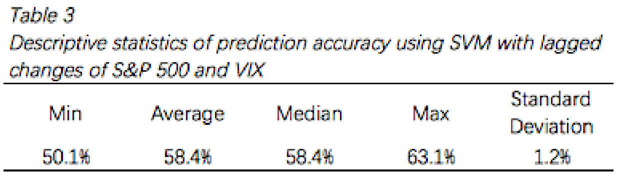


*Table 3*

*Descriptive statistics of prediction accuracy using SVM with lagged changes of S&P 500 and VIX*

| Min | Average | Median | Max | Standard Deviation |
|---|---|---|---|---|
| 50.1% | 58.4% | 58.4% | 63.1% | 1.2% |

The best accuracy(63.1%) occurs when lag is 3 and training size 650. Comparing Table 1 and Table 3, we find that adding lagged change rates of VIX as a feature improves the overall performance a little bit. This suggests past change rates of VIX does provide information about the future, but not much. Most of the predictive power is still from the past change rates of S&P 500, which has already carried most of the information provided by the past change rates of VIX because of the high negative correlation between the two predictors of the same week. Despite the high correlation in the same week, the two predictors from different weeks are almost uncorrelated. This also suggests that past change rates of VIX cannot add much predictive power to the model.

## 4 Conclusion and Furtherwork

One of the primary goals of this project is to examine the predictability of the future direction of S&P 500 Index using its lagged change rates. With Support Vector Machine, we achieved prediction accuracy of 62.7%, which suggests that the past information of the S&P 500 Index possesses significant predictive

power.
After adding past change rates of VIX as a predictor, the best accuracy increases to 63.1%. Therefore, the new-added predictor does not provide much information about the future. Again, most of the information about the future is captured by the past change rates of S&P 500.
One modification to the project would be using constant size of testing sets for different lags and sizes of training sets, instead of testing sets of different sizes. Also, same experiments can be carried out on other indices or stocks to see whether we can still achieve good performance. In addition, it would be very interesting to explore Hidden Markov model, which is trained on the past data and provides us with a probability distribution over the possible outcomes, given a sequence of states.

# References

[1] Sun Chao. *Stock Market Returns Predictability: Does Volatility Matter?*. Feb, 2008.

[2] Jingwei Chen, Ming Chen, Nan Ye. *Forecasting the Direction and Strength of Stock Market Movement.* 2013

[3] Tianxin Dai, Arpan Shah, Hongxia Zhong. *Automated Stock Trading Using Machine Learning Algorithms.* 2012

[4] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang. *Forecasting stock market movement direction with support vector machine.* 2004