Short Term Price Prediction in Financial Markets

Kam Chinta (kamchinta@gmail.com) Dharan Althuru (dharan.althuru@gmail.com) Will Yearick (yearickw@gmail.com)

1. Introduction

Asset price fluctuations have ramifications for every individual across the globe. The wealth of investors in traditional assets, such as stocks and bonds, will vary and lead to changes in personal consumption and savings. Even individuals in the most impoverished countries have their lives impacted indirectly through commodity prices such as energy and food. Foreign exchange rates will dictate purchasing power and the export / import mix of nations. The factors effecting macroeconomic trends can be attributed to a myriad of sources. Fundamental, technical, and psychological factors will all contribute to underlying price patterns and movements. Oftentimes, the cause of a certain price trajectory is unknowable and random. Clearly, professional investors, traders, and portfolio managers have their livelihood tied to price prediction. Retail investors necessarily base life-altering decisions such as college and retirement based off the whims of the stock market. Our objective is to apply Machine Learning techniques to predict short and medium term price fluctuations for financial assets. Two members of our team work directly in financial services (one as a professional trader), and all members have a keen interest in developing a better "guide" to understanding markets. Thus, we attempt to predict the price of the S&P 500 stock index over multiple horizons using twenty-two technical indicators and price data from three additional assets. We experiment with Ordinary Least Squares Regression, SMO Support Vector Regression with polynomial/ RBF kernels. The objective is two-fold: (1) to properly predict the direction of the price for 1, 5, and 10 day horizons (classification problem) (2) given a correct classification, to evaluate the delta spread of our price prediction vs the actual price subject to a given confidence band (regression problem). In order to evaluate our performance, we benchmark the algorithm vs. both a random walk and the "hit ratios" attained in prior research publications. Concretely, a hit ratio above 60% and a delta spread within 50% of the predicted price fluctuation will be viewed as a success.

2. Related work

Long-term price predictions in terms of directional accuracy on a Korean stock price index using SVMs (*Kim et al.,).* (Cao, Lijuan et al.,) performed financial forecasting to show that SVM outperform Back Propagation algorithms. In this work, we predicted long-term price prediction to start with on S&P index using SMO SVM (Platt, John et al.,) as a benchmark and results are in-line with this work. The primary goal of this project is short-term price prediction by considering data from other assets. Methods have been explored to show technical indicators improve stock predictions (Tsai et al., Agrawal et al.,). In this work, we consider technical indicators across assets for short-term directional and price predictions.

3. Dataset and Features

We obtained ten years of daily close data (source: Bloomberg) spanning from October 18, 2005 – October 20, 2015. The data pull yielded ~ 2500 potential training examples after pre-processing. Null and erroneous data points were eliminated and technical indicators which were populated for less than 80% of the data were removed as well. The assets considered were S&P 500 stock index (SPX), EURJPY foreign exchange (Euro to Japanese currency conversion), Copper future (HG1), and the Ten-year United States government bond future (TYA). The S&P 500 index is a barometer for global stock markets and one of the most liquid assets in the world. EURJPY FX tends to track global risk sentiment and oftentimes tracks the broader appetite for carry trades. Copper was chosen to represent the commodity complex and provide both an uncorrelated, technically dominated asset as well as a tracker for emerging market demand. Ten-year government bond futures represent the most liquid fixed income instrument across markets and a general flight to quality, safe haven asset.

For each of the four assets (SPX, EURJPY, HG1, TYA), we acquired daily data for twenty-two technical indicators such as Moving Average, Momentum, Hurst, etc. We also generated additional features from the raw price data including realized volatilities, exponentially weighted moving averages, and various other trend and change variables. The feature vector contains data from all four assets such as previous close price, volume, and technical indicators from previous day as well as the other generated features.

4. Methodology

We use regression techniques to predict the price and then measure the accuracy based on the directional change, essentially a classification problem. Considering previous day price as *B*, actual price as *A* and predicted price based on regression model as P, following equation is used to classify a prediction.

$$C = 1\{(A-B)*(P-B)>0\}\ where\ 1\{X\}=1\ if\ X\ is\ true;\ 1\{X\}=0\ \ otherwise \quad (1)$$

While the above equation identifies the predicted directional accuracy (PDA), the predicted price accuracy (PPA) is given below (with, $\delta = 50$). This is shown in the equation below for m predictions.

$$\sum_{i=1}^{m} \frac{1\{C[i]=1\} * 1\left\{B+(1+\delta/100)*|A-B|) > P > B+\left(1-\frac{\delta}{100}\right)*|A-B|\right\}}{\Psi}\ for\ \Psi>0$$

$$where\ \delta\ is\ the\ percentage\ spread\ and\ \Psi\ is\ number\ of\ correct\ predictions \quad (2)$$

$$\Psi = \sum_{i=1}^{m} C[i] \quad (3)$$

PPA is zero when there are no correct predictions (Ψ =0). We measure the PPE for different delta spreads.

5. Models

We perform the predictions by performing these regression approaches: (i) Ordinary Least Squares (iii) Support Vector Regression. We compare the results with *Kim et al., (2003*) and by implementing Supervised Univariate Random Walk algorithm.

5.1 Supervised Univariate Random Walk

As the name suggests, this model is only based on the target variable. Given a training set of data, a Gaussian distributed is defined with the mean and variance of the training data. For every price point, a random variable is drawn from the Gaussian distribution and it is added to the price point to make the prediction. If the horizon is k, then there are k random draws from the defined Gaussian distribution to calculate the future prediction. This is the base model to test efficiency of the other algorithms employed.

5.2 Ordinary Least Squares (OLS) Regression

To provide best short term investment guidance to the potential investor the short term market price must be predicted as accurately as possible. Since the target variable to be predicted is a positive real value, regression is an obvious tool. In the regression technique, a parameter vector is fitted to linear / polynomial features such that the predicted value is a linear combination of those features. A parameter vector is fitted to optimize the objective function of minimizing the square of the errors.

$$\hat{y}(w,x) = w_0 + w_1x_1 + \ldots + w_px_p$$

$$\min_{w} ||Xw-y||_2{}^2 + \alpha||w||_2{}^2$$

5.3 Support Vector Regression (SVR):

Financial data is essentially a time series data and support vector regression has been employed for stock market forecasting (Agrawal et al., 2013). Hence, we perform the predictions using SVR on the stock market dataset. Support Vector Regression tries to find the maximum-margin hyperplane in the higher dimensional feature space. Optimal margin is obtained by solving the following convex optimization problem:

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^Tx^{(i)}+b) \geq 1-\xi_i, \quad i=1,\ldots,m \\ & \xi_i \geq 0, \quad i=1,\ldots,m. \end{aligned}$$

ξ_i above is the "slack variable" when the dataset is not linearly separable. We use Sequential Minimimal Optimization (SMO) algorithm (Platt, John et al., 1999 and S.K.Shevade et al., 1999) using Polynomial and Radial Basis Function (RBF) kernels.

6. Experiments

Experiments are performed on a dataset having 2521 days of daily closing price, volume and technical indicator data for the four assets. We performed the experiments for (i) long term price prediction to predict for next one, two years using linear regression

and SVR to compare with Kim et al., (ii) Short term price prediction to predict for multiple horizons i.e., for next 1, 5 and 10 days and measure PPA (2) on the correct predictions. Even though the focus of this project is short-term price prediction, we performed long-term price prediction to start with to compare with Kim et al., as a baseline.

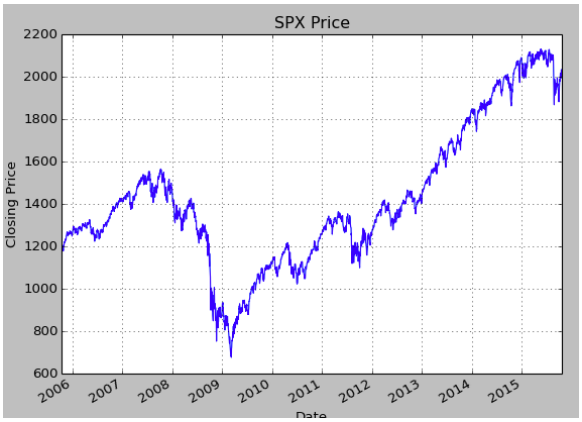

Figure1: Chart showing the closing price variation in the data

The chart above shows the closing price trend for S&P 500 Index. It is very apparent that the trend is not linear, however a portion of the curve may be fitted to a piece-wise linear regression. With this intuition, a regression model is fitted to the first degree of each one of the features, the results for the horizon periods of 1, 5 and 10 are in the order of 50, 57 and 51 percent respectively on a directional basis. However, the predicted price proximity in the being within the 50% spread is only 6, 17 and 15 percent of the results that are directionally correct. This suggests that there is a lot of room for improvement.

With more evidence for the intuition of non-linearity being established with the results of the linear regression, features of higher order are generated. Quadratic and Cubic features are programmatically generated and a feature selection criteria is established based on F-score and p-value (feature selection technique from scikit-learn is leveraged).

6.1. Short-term price prediction:

This is the primary focus of this project. Here we performed experiments with fixed window, rolling window considering features from other assets as well as technical indicators.

6.1.1 Fixed Window:

Price prediction is performed for multiple horizons i.e., for next 1, 5 and 10 days. For fixed window we considered the first 70% of training set and predicted results for next 1, 5 and 10 days. Predicted directional accuracy for next H days for this run (say R) is given by:

$$Accuracy(H,R) = \sum_{i=1}^{H} \frac{1\{C[i]=1\}}{H} \qquad (5)$$

We repeated the experiments with first 71%, then first 72% and so on till 99% i.e., for 30 runs. Accuracy for H day horizon for N runs is given by:

$$Predicted\ Directional\ Accuracy\ (PDA)\ for\ N\ runs = \sum_{i=1}^{N} \frac{Accuracy(H,i)}{N}$$
$$where\ Accuracy(H,i)\ is\ \ accuracy\ at\ horizon\ H\ for\ i^{th}\ run. \qquad (6)$$

Results

Parameter C for SVM	C=1			C=10		
PDA/Horizon	1	5	10	1	5	10
SPX	55.17	57.93	51.03	44.83	56.55	51.37
SPX_EURJPY_HG1	51.72	58.62	51.03	51.72	56.55	51.03
SPX_EURJPY_HG1_TYA	55.17	59.31	51.03	44.82	53.10	48.96

Table1:Fixed window predicted directional accuracy (PDA) for SMO SVR with C=1 and C=10

Parameter C for SVM	C=1			C=10		
PPA (δ=50)/Horizon	1	5	10	1	5	10
SPX	6.25	10.71	10.13	7.69	9.75	8.72
SPX_EURJPY_HG1	13.33	14.11	14.86	6.66	13.41	12.16
SPX_EURJPY_HG1_TYA	12.5	15.11	16.21	15.38	12.98	10.56

Table2: Fixed window predicted price accuracy (PPA) at 50% spread for SMO SVR with C=1 and C=1

We performed experiments using OLS regression, SMO support vector regression considering various combinations of features: (i) S&P 500 asset features alone (ii) features from other assets. The tables above show the results for SVR using polynomial kernel for C=1 and C=10.

For fixed window, we can see that adding features from additional assets has mixed performance considering data from all the assets (SPX_EURJPY_HG1_TYA) resulted in better performance compared to SPX feature data alone both in terms of PDA and PPA.

The following table shows the results for Polynomial Regression (quadratic features) along with a feature selection mechanism (top 25 features based on F score) in place. The accuracy measures are based on δ=50.

Horizon	**1**		**5**		**10**	
	Avg. PDA	Avg. PPA	Avg. PDA	Avg. PPA	Avg. PDA	Avg. PPA
SPX	60.00	11.10	54.60	7.30	51.60	6.40
SPX_EURJPY_HG1_TYA	40.00	33.30	49.30	43.20	50.30	36.40
SPX_EURJPY_HG1_TYA & technical indicators	56.60	5.80	46.60	8.57	46.00	8.60

Table3: Fixed Window Predicted Directional Accuracy (PDA) and Predicted Price Accuracy (PPA)

6.1.2. Rolling Window:

For rolling window, instead of considering the first 70% of training data we consider recent X% of the data to see if the predictions are time sensitive. For example, if X=10 then we considered training set as 60%-70% in chronological order and predicted for next 1, 5 and 10 day horizons. Then, we repeat the experiments considering 61%-71%, then 62%-72% and so on till 89%-99% for a total of 30 runs. Prediction accuracies are similar to the methodology explained in Section 5.1.1.

Results:

Assets/Horizon	X% (10%=1 year)	1	5	10
SPX_EURJPY_HG1_TYA (All four assets)	40%	51.72	57.24	56.89
	30%	**68.96**	**58.62**	**59.99**
	20%	41.37	51.72	52.75
SPX_EURJPY_HG1_TYA & SPX technical indicators	40%	55.17	55.86	56.55
	30%	**68.96**	**62.06**	**61.03**
	20%	34.48	51.03	52.41

Table4: Rolling window predicted directional accuracy using SVR

PPA (δ=50)/Horizon	X% (10%=1 year)	1	5	10
SPX_EURJPY_HG1_TYA (All four assets)	40%	46.66	28.91	21.81
	30%	**30.00**	**43.52**	**41.37**
	20%	25.00	22.66	22.87
SPX_EURJPY_HG1_TYA & SPX technical indicators	40%	25	30.86	28.04
	30%	**45**	**45.55**	**44.06**
	20%	50	32.43	31.57

Table5: Rolling window predicted price accuracy at 50% spread using SVR

This shows that considering around recent 30% (3 years) of training data results in better performance both in terms of directional accuracy and price predictions. Rolling window outperforms the fixed window approach in terms of prediction accuracy. This indicates the data is time-sensitive and depends on recent data for short-term predictions. We performed ablative analysis on different technical indicators and selected the indicators that yielded good performance. We observed exponential moving average, Bollinger bandwidth (BBW), relative strength index (RSI) has positive impact on both directional accuracy and price predictions. Hurst component and ROC increased PPA. Momentum, Average Directional Index (ADX) when applied on top of these decreased the accuracy and PPA. For horizon 1, best results are obtained considering data from four assets and technical indicators i.e., 68.96% accuracy and 45% PPA considering delta spread of 50.

For Polynomial Regression, the rolling window approach is slightly modified to consider the last 2% of the data to be used to fit the model. Based on experiments, this yielded a superior prediction power suggesting the time sensitivity of the features. If a larger training set is considered, it is making the model noisy and hence the fit is suboptimal. While this provides a better fit it is a non-parametric approach as it requires to constantly pull the last 2% of the data to predict next price point. Following are the results observed.

Horizon	**1**		**5**		**10**	
	Avg. PDA	Avg. PPA	Avg. PDA	Avg. PPA	Avg. PDA	Avg. PPA
SPX	43.30	15.30	54.00	29.60	51.30	27.20
SPX_EURJPY_HG1_TYA	60.00	27.70	52.60	32.90	50.60	29.60
SPX_EURJPY_HG1_TYA	**63.30**	**63.10**	**54.00**	**29.60**	**53.00**	**22.60**

Table6: Rolling Window Predicted Directional Accuracy (PDA) and Predicted Price Accuracy (PPA)

7. Conclusion and Future Work

Developing a successful asset price prediction algorithm is no small feat! Ultimately, as simple as it may sound, an asset's price will increase when there are more buyers than sellers and vice versa. Machine learning techniques, however, can play a useful role in trying to distill the vast number of impactful factors into a cohesive generative algorithm. We experimented with several models, data sets, feature vectors, and time horizons to determine an optimal approach to short term price prediction from both a classification and regression standpoint. Our results indicate that short term price prediction with SMO support vector regression performs better than linear oriented models. All models seem to perform better with a shorter training window (to a certain point!) perhaps due to efficient markets and the expedient nature of information transmission. Psychological biases and momentum factors could also be a contributing factor for the outperformance of localized vs long-term models. Initially, we believed the information from additional asset classes would likely boost performance of our primary technical indicator set. However, the introduction of additional asset classes had mixed results depending on the model and time horizon. This is likely due to shifting correlations, regime switches, and overall signal / noise deterioration.

Going forward, we have several additional avenues which we would like to explore. Towards the end of the project, we experimented with some non-linear models based off polynomial generated feature set with some encouraging results. The ability to generate additional features from the data set could provide countless future paths. We also worked on an asymmetric cost function which penalizes incorrect price direction more than accurate price **direction** prediction (absolute error squared vs absolute error). Optimizing this asymmetric cost function proved quite difficult; however, we believe the general concept could be quite interesting with more resources. Lastly, we chose four distinct assets in order to limit the scope of our project. Going forward, we could attempt to add more assets to our data set and then perform ablative analysis on the impact.

8. Bibliography

1. Kim, Kyoung-jae. "Financial time series forecasting using support vector machines." Neurocomputing 55.1 (2003): 307-319

2. Agrawal, J. G., V. S. Chourasia, and A. K. Mittra. "State-of-the-art in stock prediction techniques." International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2.4 (2013): 1360-1366.

3. Cao, Lijuan, and Francis EH Tay. "Financial forecasting using support vector machines." Neural Computing & Applications 10.2 (2001): 184-192.

4. S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy: Improvements to the SMO Algorithm for SVM Regression. In: IEEE Transactions on Neural Networks, 1999.

5. Tsai, C. F., and S. P. Wang. "Stock price forecasting by hybrid machine learning techniques." *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. No. 755. 2009.

6. Platt, John. "Fast training of support vector machines using sequential minimal optimization." Advances in kernel methods—support vector learning 3 (1999).