Tree structures for predicting stock price behaviour

Robert A. Pearson*

(Received 8 August 2003; revised 5 January 2004)

Abstract

It is shown that regression trees can be used to give useful predictions of the average price movements of individual stocks when the market is regular. While the detailed error estimates may be up to three times greater for a two month prediction than for a one week average they are still less than those obtained assuming a constant price. More qualitative measures, such as the agreement in direction of movement, and local turning points are relatively independent of the period. When it is known, a posteriori, that the market has had a minor correction the model fails. This is consistent with the chaotic, fractal behaviour. With the minor correction that occurred on the ASX during April 2000 the model actually performed better in the qualitative measures than a momentum assumption.

*School of Information Technology and Electrical Engineering UNSW, ADFA Campus, Canberra, ACT. Australia. Also 8 Sculptor St., Giralang, ACT. Australia. mailto:rpearson@netspeed.com.au

See http://anziamj.austms.org.au/V45/CTAC2003/Pea2/home.html for this article, Published August 31, 2004. ISSN 1446-8735

Contents

1 Introduction

Analysis and prediction of the stock market behaviour have been accompanied by predictions of the behaviour of the prices. Some of the approaches rely on charts of the prices, and volumes, and visual human analysis of these diagrammatic representations to suggest future behaviour. Others manipulate the historical values of the time series to calculate technical indicators. The value, or values, of one or more of these are used to suggest good times for buying or selling stock [1, 3]. Both Chartist techniques and the use of indicators are technical models which use only information gained through the trading history of a stock. In contrast a fundamental model looks at the

past financial performance of a company, the behaviour of the economy as a whole, and the industry to which a company belongs. Some also use a knowledge of the past performance of the directors in predicting the future performance. Other models mix both technical and fundamental aspects.

Recent work suggests that the behaviour of the stock market is chaotic and the time series of prices is consistent with a fractal distribution [5]. The fractal market theory can be used to derive a predictive model. In this the functional form can be considered as a particular member of the group of autoregressive functions [5]. Another approach uses past historical performance to fit a specified function of time and then extrapolates that function forward to obtain a prediction [8]. The function can include spikes corresponding to crashes [8]

The functional forms can be derived with statistical or machine learning techniques. Many machine learning models use neural networks as a tool to derive the model. Neural networks have been used for both technical [6] and fundamental models. As, theoretically , a feed forward neural network can learn any function, they could include, with appropriate inputs, any of the statistical or fractal functional forms.

Other machine learning techniques, besides feed forward neural networks, are also universal function learners. The technique of using boosted regression trees is another function learner [2]. These have been applied to derive a technical model of the stock market [4]. As with the earlier paper, the aim is to derive a model that will predict the price of a stock.

The earlier paper considered both the learning accuracy and the model predictive accuracy. The learning accuracy was evaluated by selecting a random subset of the whole data set and evaluating the errors determined by the regression trees. The predictive accuracy was evaluated by learning on historical data, predicting on the next time period and comparing this to the actual values. This paper only considers the predictive accuracy.

This paper is an extension of previous work. There are a number of dif-

ferences between this and the earlier analysis. As with the earlier work the relative change in price was predicted. One difference is that the maximum period used for prediction is increased, up to two months. One main difference is that while the earlier data was essentially a bull market this data includes a minor correction. The externally induced 'spike' in August 2001 was not considered.

2 Data and transformations

The basic data are the daily trading summaries from the Australian Stock Exchange (ASX). These are available after the close of each day's trading and most historical data from September 1998 are easily obtained. Unfortunately the data for both March 2002, and October 2002 were not provided by the data supplier. As long term averages were included in the predictions the bear market cannot be adequately tested. While over three thousand stocks are listed on the ASX some are infrequently traded. Only stock that have been listed for more than 25 weeks were considered. The average volume and average value traded per day over each week were found. Only stock where the average number traded was greater than 500 and the average value greater than $5,000 were used. When used to predict, or test the stock, the minimum average values must occur for each of the twenty five weeks before the last value. With the different form of filter used in this paper, the results differ from those for similar periods in the earlier one. The main reason for this is that fewer stock pass the newer filter than the previous one (vis, approx 500 and 800 respectively).

For each stock the averages over a number of periods were calculated. While the number of days in a calendar month vary, this analysis assumes that all months have 20 trading days. There are also 125 trading days in six months and 255 in a year. The relative changes in the periods of one week, two weeks, one month, and two months were used for the dependent variables

learning and testing. Similar averages were also used in the historical data as input (dependent variables). Another approach, to assess the behaviour of the stock was to use a quadratic least squares fit to past two weeks and *future* two weeks data.

The independent variables included the latest relative changes of the mean for longer time periods, 2, 3, 4 months, and a year. Previous historical relative changes for the previous 5 weeks and 5 months were included. As well as relative changes, a least squares curve fitting was applied to each stock. The relative slopes, the quadratic, and cubic, and fourth order terms for the closing prices were included in the independent variables. The time scales for this least squares included those of the relative changes plus the total over all the available historical data for that stock. For the volume, only the month and year periods of the least squares fitting were used. In addition the independent variables included two local variables that estimate the variation of the prices. These are the RMS and maximum deviation of the price from the best least squares linear fit. As well as these variables, a large variety of technical indicators were calculated [1, 3]. These are normalised so that all values for all stocks are similar. Where differences in values were used the normalisation was the range over the last month. Where the final values were proportional to prices or volumes the appropriate mean was used.

Some additional input variables relating to the chaotic behaviour of the stock were calculated. These were the one month, and four month range and *V-statistic*. In addition a linear least square fit was applied to the logarithm of the range and the logarithm of the period over which it was evaluated. Not all stocks have the same value of this Hurst statistic, A similar fit was also applied to the *V-statistic*. Brownian not chaotic motion would have a constant value of the V-statistic. This V-statistic shows that stock price movements are not Brownian [4]

Not all the available stocks and dates were used in training. Earlier trading weeks were selected through a pseudo random process where the probability of selection decreased as the time between the current date and

the training date increased. This was similar to the process used in the earlier paper. This sample was partitioned into two. The one used to construct the trees contained over 4,000 examples. The other used to select the pruned subtree and stop the boosting had over 2,000 examples. In this paper five different periods in 2000 are used for testing. For four, the last day of the average for the two month prediction corresponds to the end of a quarter. The other corresponds to a minor *correction*, and decrease in the price of certain stocks, with the period for prediction beginning on 14th March.

3 Results

3.1 The momentum assumption

The regression trees learn the difference between a naive prediction and the actual. The naive prediction was chosen to be related to a constant increase or decrease in price. The factor chosen in the previous paper was 0.75 , while some earlier tests used 0.5 . This factor was selected on the basis of the learning accuracy and observations on the output. It was observed that no single value was the best for all error estimates. Nor was a single value at a particular date the best for all time periods. Also for a given period, the best value was not the same for different dates of the testing. As in the previous paper, this one uses 0.75 as a compromise value. For the quadratic term in the least squares fit, the naive assumption was a zero value.

3.2 Predictive behaviour

When the errors over all the stocks are evaluated, the model has a considerable advantage when the market is fairly regular. The results for the errors at the end of the four standard quarters are similar. In all these relatively

TABLE 1: Errors for predictions at different times

	Mean			RMS			LAD			Maximum		
	Model	Naive	Same	Model	Naive	Same	Model	Naive	Same	Model	Naive	Same
February 5												
one week	-0.07	-0.17	1.1	-0.17	1.0	0.94	0.55	0.63	0.57	10.8	7.1	7.4
two weeks	0.02	-0.10	1.12	-0.17	1.03	1.01	0.51	0.64	0.60	6.6	7.2	7.8
one month	0.10	-0.11	1.33	-0.27	1.43	1.37	0.80	0.82	0.76	10.3	13.8	14.1
two months	0.26	0.06	1.24	-0.48	2.32	2.02	1.32	1.26	1.09	21.3	19.6	19
April 14												
one week	1.15	0.93	1.1	2.15	1.4	1.7	0.4	1.01	1.18	8.9	4.7	4.7
two weeks	0.89	0.83	1.12	1.74	1.43	1.79	0.4	1	1.24	12.3	6.3	5.0
one month	1.22	1.45	1.33	2.01	2.43	2.16	0.42	1.64	1.5	8.7	10.7	5.9
two months	2.16	1.60	1.24	6.05	3.32	2.13	0.56	2.09	1.5	43	21.1	6.1
quadratic	0.49	n/a	0.32	0.81	n/a	0.58	0.53	n/a	0.40	2.7	n/a	2.2
November/December												
one week	-0.05	-0.13	-0.1	0.65	0.82	0.73	0.4	0.48	0.43	4.4	6.7	5.6
two weeks	0.01	-0.12	-0.06	0.62	0.82	0.77	0.4	0.51	0.46	4.0	5.4	5.1
one month	0.01	-0.20	0.01	0.63	0.9	0.95	0.42	0.59	0.57	4.6	5.4	7.3
two months	0.02	-0.13	0.17	0.80	1.08	1.29	0.56	0.75	0.78	3.9	7.5	9.2
quadratic	-0.08	n/a	-0.11	0.34	n/a	0.33	.21	n/a	0.20	2.5	n/a	2.4

TABLE 2: Direction of price Movement

Prediction	Zero	Increasing				Decreasing			
	Actual	Actual	Model	Model	Naive	Actual	Model	Model	Naive
February 4									
one week	15	256	285	66%	52%	249	219	70%	51%
two weeks	22	246	236	72%	61%	252	264	70%	59%
one month	16	245	262	69%	65%	259	246	72%	66%
two months	3	262	259	76%	65%	255	259	75%	70%
April 14									
one week	8	63	228	21%	21%	178	218	93%	92%
two weeks	12	89	156	38%	30%	194	179	91%	85%
one month	6	92	186	38%	21%	197	201	93%	83%
two months	8	104	210	29%	19%	221	218	83%	77%
November/December									
one week	17	224	188	71%	62%	178	218	56%	50%
two weeks	16	209	220	65%	54%	194	179	64%	52%
one month	13	209	218	74%	62%	197	201	73%	54%
two months	14	184	196	68%	60%	221	218	75%	65%
quadratic		255	221	52%	n/a	129	198	38%	n/a

normal situations, the actual error estimates of the model are less than the alternatives. For the period corresponding to the minor correction of the stock market in April, the model does not perform well (Table 1). During this period the assumption of constant prices has the lowest error. Attempting to predict the actual value of the quadratic term directly is neither significantly accurate in the decrease in error during learning nor is it useful.

Trading decisions can use the anticipated direction of the movement of the stock price. The values for some of the quarters are included in Table 2. For both the model and the naive prediction, values for the percentage correct given the forecast, are included. Note that some stocks have the same average value in the next period. The model usually performs much better than the naive or a simple uniform distribution. Note that during the correction a

TABLE 3: Turning points

	Maximum			Minimum		
	Actual	Model	Both	Actual	Model	Both
February 4						
one week	66	39	25	109	77	57
two weeks	75	54	31	80	49	37
one month	83	55	32	64	20	16
two months	92	83	46	56	46	30
April 14						
one week	107	51	41	15	98	8
two weeks	93	49	43	31	50	11
one month	193	120	111	27	44	13
two months	196	69	57	50	44	20
November						
one week	47	90	34	70	34	25
two weeks	52	27	13	88	37	31
one month	41	36	23	87	82	55
two months	53	27	18	68	41	30

number of stocks actually increased in price and a very large number of stocks decreased in value. Both the model and the naive assumption give a prediction where more stocks rose in value.

As the model predicts the new value, it can also predict a turning point. Neither the constant price assumption, nor any simple variations on the momentum assumption can give this behaviour. Many stocks, of course, continue rising or falling, so the numbers of estimates of minima and maxima are much less than the total number considered. The values for all periods are included in Table 3. During the minor correction some stocks had a local minimum. As expected with a correction, the number of local maxima are significantly more than the minima. The model tends to give less than the correct number of maxima and more of the minima. In a more regular market

the model is more likely to predict a lower number of extreme values than actually occur. It is also likely to predict turning points which do not occur.

4 Discussion

4.1 Error behaviour

For August/September the mean error of the two month prediction (−0.08) is actually less than that of the one month prediction. For the other error estimates, the two month prediction is the least accurate. For June all error estimates increase with estimation period. The pattern in January/February and November/December is more mixed. For February/March the two month prediction is on average much greater than for the other periods, but the RMS is significantly less than the naive, or constant price prediction. While the pattern of the error between different periods for predictions differs with the date chosen, the shorter period predictions are usually more accurate than the longer ones.

4.2 Direction of price movement

While the errors tend to increase with the length of the period for the prediction, the agreement between the direction predicted and the actual price movement is relatively unchanged. The percentage correct given a forecast can actually be the largest for the two month prediction. In some cases the percentage correct given a forecast is close to that where a coin toss occurs for each case. In others most forecasts are correct. In all cases the model performs better than a momentum assumption. The other consideration is how many cases are missed. The worst case of the regular was 41 percent for the one week December prediction of increasing price, the best was 12 for the one

week decreasing prices in September. During the correction the values varied between 23 for the one month decreasing and 43 for the one week decreasing. The momentum assumption missed between 30 (for two weeks decreasing) and 53 (for one month increasing). The model was actually better than a momentum assumption in this minor correction.

The agreement between the predicted local turning points and the actual ones is also relatively constant for the different time periods.

4.3 Chaotic behaviour

When there is a sharp change, or correction, the fractal market hypothesis suggests that a useful forecast is unlikely. The chaotic nature of the market also suggests that at certain times a machine learning technique will be appropriate but fail at other times [7]. The minor correction in April is expected (a posterior) to be one of the times when the model fails. This particular period for the Australian market was not a complete correction. Some stocks continued to increase, some actually had a local minimum (that is, they increased after previously decreasing). The model predictions for one month actually had significantly more maxima than minima. All the predictions suggested significantly fewer stocks would be at local maxima than the actual case.

The a posteriori examination of the errors suggests that the other dates considered were fairly regular. The June quarter contains more local maxima at the two month time scale than the others. Similarly the September quarter at the one month scale has many more maxima than minima. The model in this case also predicts a significant asymmetry. During the correction, the model at the one month time scale predicted an asymmetry, but not nearly as large as that which occurred. The one week predictions suggested that nearly twice as many minima than maxima would occur. The reverse occurred.

Overall, the model tends to give useful predictions when the stock market

is fairly regular. The contrast to this regular behavior is a *joker* or a change of local strange attractor. Unfortunately the *regularity* or *joker*, can so far only be estimated a posteriori.

4.4 Extra variables

This revised model includes some variables not considered in the earlier one. Some of these appear as important variables. The Hurst statistic is the second most important variable for the one month predictions in May/June. The 80 day V-statistic appears at rank 8 for the November/December 2 month predictions. The 20 day V-statistic for price, and the 20 and 40 day V-statistic for volume were also in the importance lists. The error estimates between the actual prices and the constant momentum values also appeared sometimes. The Hurst value for volume did not appear, nor did the gradients of the V-statistics. The error estimates between the actual prices and the constant momentum values also appeared sometimes.

4.5 Model application

The regression trees are *on average* learners. Selecting only a few stocks is not likely to be a rewarding trading strategy.[1] As well as model predictions a useful trading strategy both for buying and selling is needed. Also some techniques for capital management must be included in the strategy. Another requirement is a large enough bank balance to spread across a number of stocks that are selected with the strategy. The predictions also assume that knowledge of the predictions will not influence the market. The trading using the model must be of insignificant volume.

[1]For interest I tried this in a stock prediction competition. This had a limited portfolio value of $100,000 for each month. I selected some stocks with the predicted largest gains. Most months I lost. In one I was in the top ten. On average over the all the months I was about even. Successful stock pickers had much more success and considerable gains.

Even with a useful trading strategy, and suitable bank balance, the model is known to fail. It is a technical model using data from trading. It does not included, except a posteriori, decisions by individual firms, nor economic influences such as interest rates. It cannot predict either *crashes* or the consequences of *terrorist* attacks, or wars.

5 Conclusions

The relative average change of price can be predicted by boosted regression trees when the market is regular. The magnitude of the errors tends to increase with period. Behaviour of prices, such as direction of movement, local maxima and minima, can also be usefully predicted by the output. These more qualitative predictions are relatively independent of the period over which the average change is to be predicted.

The regression tree model cannot predict actual prices usefully when the market has a major change in behaviour. Any major correction corresponds to a *joker* or a change in local strange attractor. As anticipated this model, as any technical model, fails in such situations. While the model fails to adequately consider even a minor correction the qualitative measures of success are better than those of either a constant price or a momentum assumption.

Acknowledgments: Calculations were performed at UNSW ADFA.

References

[1] S. B. Achelis. *Technical Analysis from A to Z.* 1992. C951, C954

[2] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistis (submitted)*, 1999. C952

[3] T. A . Myers. *The Encyclopedia of Technical Market Indicators*. 1992. C951, C954

[4] Robert A. Pearson. How to gain?/lose? on the stock market - datamining the ASX. In Vishy Karri and Michael Negevitsky, editors, *AISAT'2000*, pages 237–242, Hobart Australia, December 2000. C952, C954

[5] Edgar E. Peters. *Fractal Market Analysis : Applying Chaos Theory To Investment And Economics*. J. Wiley and Sons, New York, 1994. C952

[6] Zhang Ruying, Guan_Seng Khoo, and Lawrence Ma. Devising a trading strategy based on the forecast slopes of time series using a neural network. In *ICONIP'99 6th International Conference on Neural Information Processing*, pages 1123–1126, Perth Western Australia, November 16-20 1999. C952

[7] L. A. Smith. Sane, phychic and physchotic neural networks. In *European Geophysical Society Newsletter: European Geophysical Society XXIV General Assembly*, Le Hague, April 1999. C960

[8] Wei Xing Zhou and Didier Sornette. The us 2000-2002 market descent: How much longer and deeper? *Quantitative Finance*, 2:268–481, 2002. C952