"Forecasting the 30-year U.S. Treasury Bond with a System of Neural Networks"

Wei Cheng, Lorry Wagner, and Chien-Hua Lin

(c) Copyright 1996, Finance & Technology Publishing, Inc., P.O. Box 764, Haymarket, VA 20168, USA. All rights reserved.
Published in the NeuroVe$t Journal, January/February 1996.

ABSTRACT

A forecasting model based on a system of artificial neural networks (ANN) is used to predict the direction of the 30-Year U.S. Treasury Bond on a weekly basis. At the close of Friday's market, the 23-variable database is updated with the latest information and the data pre-processed for input to the prediction system. Thirty-two feed-forward neural networks are trained on the new data and then individually recalled to predict the following Friday's market direction. These results are then used as input to a decision model that ultimately determines the final prediction. Forecasting began in 1989, with live trades commencing in 1990. The average accuracy of the buy prediction was 67% over a five-year period, with an average annualized return on investment (ROI) of 17.3%. This compares with an ROI of 13.9% for the Lehman Brothers T-Bond Index. This paper describes the methods used for data selection, training and testing, the basic system architecture, and how the decision model improved the total system accuracy as compared to individual networks.

1 INTRODUCTION

The two characteristics of a successful financial forecasting system are the exploitation of inefficiencies of a given market and the precise application to that market. Overwhelming evidence now indicates that opportunities exist for consistent positive returns over a given period of time, in contrast to the traditional Efficient Market Hypothesis. One need only read Market Wizards (Schwager [1992]) or The Reminiscences of A Stock Operator (Lefevre [1985]) to see how true this is. However, it is also clear from studying these accounts that successful traders invariably apply two rules to their forecasting methods. No matter how good their system is, it must be applied unemotionally and consistently. Whenever ego or rational judgment enters into the trading, they have lost money. They have also recognized that every winning system will have losses and that these losses have to be accepted and minimized as much as possible. Very simply, one must see the market as either verifying a trade or moving against it. Once the focus turns to profit and loss, objectivity, as well as money, will be lost.

The goal of this research effort was to develop an accurate neural network forecasting system and apply it as described above. The first part of developing a methodology focused on eliminating, as much as possible, business risk. Compared with stocks, government treasury securities have minimal business risk since they are backed by the U.S. Government. Immunity from default also makes global trading easier than with most other investments. A second emphasis was placed on using a security that appeared to be amenable to fundamental analysis. Again, this led to the U.S. Treasury market, the largest financial market in the world, with over 3 trillion dollars in securities traded on an around-the-clock basis. The 30-Year U.S. Treasury Bond was selected because it is a significant variable in many econometric and financial models, it yields the greatest return on investment, and it is a highly-liquid asset.

The third component of this research focused on finding an improved technique for establishing trading positions. Artificial neural networks (ANNs) appeared to have this potential. Significant research using an ANN approach had been done in a variety of financial areas. In the bond-rating model of Duddy & Shekjar [1988], ANNs were compared to standard regression techniques, where accuracies of 88.3% versus 64.7% were achieved, respectively. Odom and Sharda [1990] looked at bankruptcy prediction using discriminant analysis and neural networks. Their results gave the edge to the ANN by a 77.8% to 70.4% margin. In another instance, involving stock price pattern recognition, Kamijo & Tanigawa [1990] developed an ANN which recognized the correct pattern in 15 of 16 cases studied (93.8% correct). The forecasting system of Kimoto and Asakawa [1990] was able to achieve an 18.6% greater profit trading the TOPIX index on the Tokyo Exchange compared with a traditional buy and hold strategy. Finally, Chuah [1992] reported forecasting the New York Stock Exchange end-of-month and end-of-quarter stock price indices using an ANN that yielded a 116% return on investment (ROI), compared to 94% using a buy-and-hold strategy, and 38% using a linear model.

This paper illustrates a successful bond-forecasting system based on ANN's. The final trading model was developed to predict the long-bond market as either an up movement or a down movement. The system forecasted market trends on a weekly basis, i.e., based on the closing price (and yield) for a given Friday, it predicted the direction for the following Friday. Each week, the results from 32 individual networks were used as input to a final decision model. The networks' performance were measured by their accuracy of classification and the decision model was rated according to its ROI.

2 DATA SELECTION AND PREPARATION

The selection of data and its preparation for modeling represent crucial tasks in the development of any forecasting model. These tasks have been addressed in Caldwell [1994] and a special issue of the NEUROVE$T JOURNAL [1995]. The phases of data preparation can be broadly classified into three distinct areas: variable selection and collection, data inspection, and data pre-processing.

Treasury issues, especially the long bond, are impacted by the Federal Reserve's monetary policy as well as U.S. and world economic conditions. There are literally hundreds of variables that provide measurements of such policy and conditions. Since our goal was to trade on a weekly basis, we were able to narrow the list of choices to those variables relevant to that time frame. The data were initially obtained from two sources: interviews with successful traders and the indicators used by the investment firm supporting this research. A list of variables was compiled and separated into the following three categories:

Group I — data consistently used by the traders and the firm.
Group II — data consistently used only by the traders.
Group III — data consistently used only by the firm.

The data were then collected over the period 1977 through 1988. The Group I variables (14 in all) were selected as the starting point for network development.

It is imperative that all data (especially financial) be cross-checked in as many ways as possible to ensure accuracy. Even though the data were obtained from reliable sources (i.e., Bloomberg News Service, Telerate, The Wall Street Journal and Barron's), significant errors were discovered. The validity of the input data was checked both visually and statistically, by observing plots of relevant measures (mean, maximum, minimum, variance, standard deviation, frequency distribution, outliers, etc.) of the data, and cross-checking between the multiple data sources. Most of the extreme data values were not in error, i.e., they usually represented random or unusual events, such as the stock market crash in October 1987). We chose to include all data for each variable, even though many traders discounted such events as being unrelated to long-term bond movements.

The final preparation of the data was to scale each variable to values between 0.1 and 0.9. This provided a method for normalizing the data across common ranges, which allowed the networks to map the data without forcing the transfer functions to their limits. The input values (as well as the desired output values) were normalized based upon the extreme values for each variable. For each variable with a maximum value, Mi, and minimum value, mi, the scaling was applied as follows:

$$Vi = (Vi' - 0.9*mi) / (1.1*Mi - 0.9*mi)$$

where
Vi' is the raw data
Vi is the scaled inputs

The ten percent buffer above and below the scaled range was provided to compensate for cases where future values might be higher or lower than the initial limits.

3 MODEL DEVELOPMENT

The goal at this early stage was simply to develop any network (using the 14 variables from Group I) that could forecast with better than 50% accuracy for at least 30 weeks on data it had not seen before. Initial network design began using back-propagation networks. Other models were then designed using counter-propagation, probabilistic, and functional-link networks. The basic models were trained by presenting from one to ten weeks of past data to the input layer and then adding neurons to a single hidden layer (starting with one neuron). Although several networks forecasted consistently well (with 57-59% accuracy), the best single network (with 61% accuracy) was used as the benchmark for all future work. This back-propagation network used ten weeks of input data (140 processing elements), two types of hidden layers (one with the sine transfer function and one with the sigmoid), each with 5 elements, and an output layer of one element.

The next step involved the process of individually adding the variables from Groups II and III. When the back-tested performance surpassed that of the benchmark network, the new variable was added to the list. The results from applying this procedure produced the 23 variables listed in Table 1.

The next phase of the project involved a systematic exploration of training time, network structure, data set size, and randomization methods. Whenever modification to any of these led to a network with superior performance over the best current network, the previous network was replaced with the new one. In the exploration of training iterations (where each iteration represents one input vector presented to the network), the length of training each network was determined empirically. Networks were trained and recalled on the test set at intervals from 500 to 100,000 presentations of data. These occurred in steps of 500 (up to 10,000), steps of 5000 (up to 25,000), and steps of 25,000 (up to 100,000). Typically, at 2000 iterations, the networks had converged, however at 5000 learns (with roughly the same mean squared error based on the training set) the forecasting performance on the test set was superior. Continued training produced improved training errors but poorer forecasting results. Therefore, 5000 training iterations were used for all of the networks in the live forecasting model.

As part of the investigation of network size, the 90/10 cross-validation method was selected as the best use of resources. This test employed ten copies (with different random weights) of each architecture, with each network trained on 90% and tested on 10% of the data (from 1977 to 1986). This procedure was repeated 10 times to allow for testing the networks over all of the data.

The input space for the variables in Table 1 was explored over periods of 10 to 26 weeks (in steps of 10, 15, 20, and 26), where the size of the input vector was 23 for one week, 46 for two weeks, etc. The hidden layer was varied from 1 to 100 neurons (in steps of 1, 5, 10, 15, 20, 30, 40, 50, and 100) using the sine or sigmoid transfer functions, while the output layer was varied from 1 to 2 neurons using a linear transfer function. Outputs of price and yield were applied to all combinations. Final results were based upon the 10 average results for each of the architectures (a total of 3840 networks). Results indicated that networks using sine transfer functions in the hidden layer, input spaces of 20 and 26 weeks, and hidden layers with 20 to 40 neurons performed best. Networks with 1 and 2 output neurons provided equivalent predictions for the first week. However, results for the second week were unreliable, with prediction accuracy ranging from 45% to 55 % over the 10 test sets.

Further accuracy improvements were achieved by varying the size of the training data window. Up to this point, each of the networks had been trained using all of the available data (520 weeks covering the period 1977 through 1986). A systematic search using input spaces of progressively smaller training sets discovered maximum performance with a training window size of 100-200 weeks. Although this paring of the data set produced greater forecasting accuracy, it seemed unusual to ignore valid data. However, all of the bond traders interviewed as part of this study also chose to ignore past data beyond one or two years.

Initially, network weights were re-randomized prior to re-training using new weekly data. However, an alternative method, which utilized the weights from the network trained in the previous week as the initial weights for newly-trained networks, proved to be significantly superior. This procedure was run in parallel with one where networks weights were always re-randomized prior to re-training. After providing superior performance over the period of 1990, the alternative method was adopted for tests performed on data beginning in 1991. Network accuracy over the 8/90 to 12/90 period was 66% compared to the live model results of 52%.

Ultimately, 32 networks with forecasting accuracies greater than 55% were selected. Potentially, all of these networks were profitable. The challenge, of course, was to discover those that would provide the best forecast and yield a positive ROI. Another consideration important in trading systems is the maximum consecutive loss, i.e., the greatest amount of money one loses over successive weeks. The best network in terms of accuracy was usually not the one with the lowest maximum consecutive loss. Therefore, in order to achieve the overall goal of maximum ROI with minimum risk, we considered the development of two separate networks. The concept was to combine networks which were the most accurate with ones which were least risky.

Table 2 summarizes the structure of the 32 networks. Each network has three layers: one input layer, one hidden layer and one output layer (see Figure 1). Their structure consists of the following: 16 different networks predict the bond yield and 16 predict the bond price. The hidden layer has 20, 30, 40, or 100 processing elements; the output layer has 1 or 2 elements used to predict 1 or 2 weeks into the future; and the number of weeks of historical data presented to the input layer are in series of 20- and 26-week training vectors.

4 INDIVIDUAL NETWORK FORECASTING

Research results demonstrated that individual networks were able to predict bond direction much better than the magnitude of the price variance. This indicated that the networks were capable of matching the frequency and the phase of the actual bond signal (yield and price), but not the absolute magnitude. Figure 2 illustrates this finding. For week seven for example, the network predicted the down market correctly. However, it did not accurately predict the magnitude. It predicted a market price drop of about 1.5% (from 105.25 to 104.75), when the actual bond price dropped more than 3.5% (from 103.75 to 100.25).

Although the difference in the structure of the networks and their individual results over the invest-ment period varied widely, 29 of the 32 networks appeared to contribute to the forecasts. Figure 3 summarizes the contribution of each individual network to the final weekly prediction. It was clear that nine networks were predominant in creating the final forecast. However, another 20 networks contributed to achieving greater ROI and lower maximum consecutive loss than any single network accomplished individually. Furthermore, different combinations of networks contributed to weekly forecasts. Thus, even if one were tempted to choose the top 10 networks based on contributions, the same results could not be achieved.

5 COMBINED NETWORK FORECASTING

Various combination techniques have been proposed by different researchers. Rogova's [1994] methods improved classification accuracy while producing a 15-30% reduction of misclassification error compared with a best individual classifier. Lincoln and Skrzypek [1990] combined five back-propagation networks by averaging the outputs and assigning an adjustable weight to each output. In other research, Shi and Liu [1993] developed networks which combined outputs from three different models: ARIMA, trend analysis, and Brown exponential smoothing. The combined results were much better than those for any of the individuals ones studied.

Lincoln and Skrzypek combined networks in a convex combination as follows:

$$\acute{Y} = \text{sum}(wk*yk) \qquad (k=1, 2, ..., n)$$

where

Ý is the cluster output value

yk is the output value for network k

wk is the weight assigned to network k's output value.

and

wk is adjusted to take into account the recent reliability of the network. The weight adjustment rule was wk = wk*G(e/ek) where:

$$e = 1/N*\text{sum}(ek) \qquad (k=1,2,...,n)$$

G is the gain of the adjustment

ek = ||Ý - yk|| which is the network deviation from the cluster output.

As discussed previously, the networks in the present model were designed to predict market direction rather than magnitude. Therefore, it was meaningless to use the output value yk (and likewise wk) in the model. Rather than use the exact network output values, a mapping model was developed as follows:

First, calculate y'i for each network's output yi based on:

$$y'i = \begin{cases} 1 & \text{if yi predicts an up market} \\ -1 & \text{if yi predicts a down market} \end{cases} \qquad (i = 1, 2, ... , 32)$$

Second, calculate the percentage accuracy pi for each network based on past weeks' performance.

Third, find the top 3 highest percentages of all networks {H1, H2, H3}, and calculate a summary value Y:

$$Y = \text{sum}(y'i*p'i) \qquad (i = 1, 2, ..., 32)$$

where

$$p'i = \begin{cases} pi & \text{if pi is a subset of \{H1, H2, H3\}} \\ 0 & \text{otherwise} \end{cases}$$

Last, map Y as follows:

$$\text{Prediction} = \begin{cases} \text{BUY if } Y > 0 \\ \text{SELL if } Y < 0 \end{cases}$$

Table 3 compares the results of the combined model and the best individual network. It is clear that the decision model performed the same or better in all aspects. The combined model increased forecasting accuracy: the best network (network 22) achieved 60% accuracy while the combined model achieved 67% accuracy. In the key areas of ROI and consecutive losses, the combined model increased ROI from 11.66% to 21.02% while it decreased the maximum consecutive losses from 8.17% to 4.83%.

6 SYSTEM INVESTMENT PERFORMANCE

The investment strategy employed in the live trades used a buy strategy, i.e., enter the market only when the prediction is a buy. Table 4 illustrates the annual results based on the original investment of $10,000,000.

Table 5 presents the complete results for the performance of the trading system. The Lehman Brothers T-Bond Index was used for comparison.

With an annualized average capital gain of 10.3% (6.97% derived from coupon and money market interest), the system achieved a total ROI of 17.27% versus 13.88% for the Lehman Brothers T-Bond Index. One of the key characteristics of the system was its relatively low risk. The maximum consecutive loss for the 5-year period was 4.72% (1990), while the largest single weekly loss was 3.8% (1990). These returns and losses are actual balance sheet values, since there were no hidden transaction costs. In trading the long bond, the price quoted on a bond by a broker is the net price to the buyer.

CONCLUSIONS

The 5-year live performance results demonstrate that an ANN can be a powerful tool for forecasting the long bond. Results indicate that the system of neural networks developed here predicted the direction of the long bond more accurately than its magnitude. It was found that a network's prediction accuracy was determined not only by its structure but also by the amount of historical data used in the training set, the number of weeks in the input vector, the number of learning iterations, and how the data is pre-processed. Based on the goals of maximum ROI and minimum risk, results suggest that, in order to obtain the best prediction system, one should use multiple networks rather than only the best single network. By combining the prediction from 32 networks into a single forecast, a combined model gave better results than any single network. It appears that the individual networks extracted different information from the data. In addition, the contribution from some variables were more significant at different times over the trading period and certain networks gave more weight to these data. This led to the use of different networks for making the weekly forecasts, with a total of 29 (from an initial set of 32) networks contributing to the final results during the trading period.

It is also apparent that the system developed had a finite lifetime of successful forecasting. During the 1989 to 1993 period, it did not make an incorrect prediction more than two weeks in a row and did not drop below a 50% forecasting accuracy over any ten-week period. During the last two months of 1993 and the first month of 1994, the system did not maintain its previous accuracy. Thus, in February 1994, investments using this model were discontinued. When it became clear that the system could not accurately forecast the value of the bond (see Figure 2), two choices were available:

- redevelop from the start, or
- attempt to utilize the available forecasts in some way.

Since the predictions were accurate in terms of phase and frequency, we could have attempted to make use of this information. However, we considered that, if the data were pre-processed differently, better forecasting results of the actual bond value would probably have been obtained.

As it turned out, the training of the large networks developed as part of this research were very time consuming. For the largest networks, 16 hours on a 386/25 w/387 & 3167 co-processors and 3 hours on a 486/33 w/4167 were required. When the long bond yield reached a historical low (5.78), the entire yield curve became compressed, and the model failed to perform. It is also possible that reducing the training set to the previous two years produced a model that was exceptional in an up market, but could not perform in a down market.

Applying this technology to financial markets is still more of an art than a science. Further research should focus on developing models to answer this market question as well as developing networks that directly forecasts the bond's direction. It is also be possible that pre-processing the input data more efficiently using newer computer technology along the use of improvements to the backpropagation algorithm can provide better and faster results. Another promising area appears be to the use of genetic algorithms for developing systems of combined forecasting networks.

ACKNOWLEDGMENT

The authors would like to thank Abraham Feldesman for his continuous support during the course of this work.

REFERENCES

Caldwell, R.B. [1994] "Design of Neural Network-based Financial Forecasting Systems: Data Selection and Data Processing," NEUROVE$T JOURNAL, Vol.2, No.5, pp. 12-20.

Chuah, K. L. [1992] A Nonlinear Approach to Return Predictability in the Securities Markets Using Feedforward Neural Network, Ph.D. dissertation, Washington State University.

Duddy, S. and Shekhar, S. [July 1988] "Bond Rating: A Non-Conservative Application of Neural Networks," Proc. of the IEEE International Conference on Neural Networks", pp. 443-450.

Kamijo, K. and Tanigawa, T. [1990] "Stock Price Pattern Recognition: A Recurrent Neural Network Approach," Proc. of the IJCNN, San Diego, Ca, pp. 215-221.

Kimoto, T., and Asakawa, K. [1990] "Stock Market Prediction System with Modular Neural Networks," Proceedings of the IJCNN, San Diego, Ca, pp. 1-6.

Lefevre, E. [1985] Reminiscences of a Stock Operator, Traders Press, Inc.

Lincoln, W. and Skrzypek, J. [1990] "Synergy of Clustering Multiple Back Propagation Networks," Advanced Neural Information, 2, pp. 650-657.

NEUROVE$T JOURNAL [1995] Special issue on Data Selection and Preprocessing, Vol.3, No.6.

Rogova, G. [1994] "Combining the Results of Several Neural Network Classifiers," Neural Networks, pp. 777-781.

Shi S., & Liu B. [1993] "Nonlinear Combination of Forecasts with Neural Networks," Proceedings of the IJCNN, pp. I959-962.

Schwager, J.D. [1992] The New Market Wizards, Harper Business.

Weiss, S. and Kulikowski, C. A. [1990] Computer Systems That Learn, Morgan Kaufmann Publishing, Inc., San Mateo, CA.

Wei Cheng is a Ph.D. candidate in Business Administration with a major in Computer Information Science at the Cleveland State University, with interests in neural network forecasting for business and finance, and time series analysis and modelling. Chien-Hua Lin is currently Associate Professor and Chairman of the Computer Information Science Department at Cleveland State University, with a Ph.D. in operations research from Case Western Reserve University, Cleveland. The above authors can be reached at Cleveland State University, CIS Department, 2121 Euclid Ave. UC482, Cleveland, OH 44115 USA.

Lorry Wagner is currently CEO for Phillips Electric. During the course of the neural network research, he was Director of Research for Global Asset Management and directed the work described in this paper. He received his Ph.D. at Purdue University. He can be reached at Philllips Electric, 4126 St. Clair Avenue, Cleveland, OH 44103.

YIELD CURVE DATA:
Federal Funds Rate
3-Month Treasury Bill Rate
6-Month Treasury Bill Rate
1-Year Treasury Bill Yield
2-Year Treasury Note Yield
7-Year Treasury Note Yield
10-Year Treasury Bond Yield
30-Year Treasury Bond Yield
30-Year Treasury Bond Price

INDICES:
S&P 500
Dow Jones Industrial Average
Dow Jones Utilities Average
Change in M2 from Last Week
Buy or Sell Indicator (1= Buy, 2 = Sell)
DLJ Index(Treasury - Junk Bond Spread)

FOREIGN RATES:
Euro Dollar Rate
Dollar/Yen Ratio
Dollar/DM Ratio
Japanese 10-Year Bond Yield
German 10-Year Bond Yield

COMMODITIES:
Cash Price of Crude Oil
Cash Price of Gold
CRB Index

TABLE 1. The 23 variables which surpassed benchmark tests.

Network Task	Hidden Nodes	Output Nodes	Input	Number of Networks Tested
Yield	20, 30, 40, 100	1, 2	20-week (460) 26-week (598)	4*2*2=16
Price	20, 30, 40, 100	1, 2	20-week (460) 26-week (598)	4*2*2=16

TABLE 2. Structure of the 32 networks selected.

	Combined Model Results	Best Individual Network (#22)
Accuracy	67%	60%
ROI	21.02%	11.66%
Maximum Loss	(2.76)%	(2.76)%
Maximum Consecutive Loss	(4.83)%	(8.17)%

TABLE 3. Comparison of combined model and best individual network.

	1989	1990	1991	1992	1993
Basis Points Made	115	85	105	91	52
Capital Gain	$1,322,500	$977,500	$1,207,500	$1,046,500	$598,000
Coupon + MM	$849,404	$824,154	$723,628	$644,274	$436,549
Total Return	$2,171,904	$1,801,654	$1,931,128	$1,690,774	$1,034,549

TABLE 4. Annual results of systems tested.

	1989	1990	1991	1992	1993	Average
Annualized Capital Gain	13.23%	9.78%	12.08%	10.43%	5.98%%	10.30%
Coupon and MM Interest	8.49%	8.24%	7.24%	6.44%	4.46%	6.97%
Annualized Total Return	21.72%	18.02%	19.32%	16.87%	10.44%	17.27%
Lehman Brothers Index	18.68%	7.57%	16.98%	8.86%	17.31%	13.88%
Maximum 1 Week Draw Down	(1.61%)	(3.80%)	(2.42%)	(1.50%)	(2.98%)	(2.46%)
Consecutive Weeks Draw Down	(1.61%)	(4.72%)	(2.42%)	(1.73%)	(2.98%)	(2.69%)
Number of Buys	25	26	27	37	33	30
Prediction Accuracy	68%	54%	81%	65%	64%	67%

TABLE 5. Final results for the final trading system compared with the Lehman Brothers T-Bond Index.