Capstone Project: Final Teleprompter Script

**Can Neural Networks outperform Linear Regression in Stock Market Prediction?**

*Mohamed Mohamar*

*Data Science Institute, Saint Peter's University*

[*mmohamar@saintpeters.edu*](mailto:mmohamar@saintpeters.edu)

**Contribution of Competitor’s Article**

Good evening everyone, my name is Mohamed Mohamar. My presentation will not be more than fifteen minutes. Here is how I will proceed: I will first talk about my competitor’s article; then present you my novel contribution; my data, where it came from and its content; I will talk about my method; I will present some of my quantitative results; I will compare my results with my competitor’s results; I will have discussions about my results; I will have a discussion on questions performance on big data and time measurements; I will finish with my conclusions and take few questions. Let’s get started.

A lot of great work has been done in the field of Data Science using Neural Networks, applied to varieties of disciplines of humanity. I choose this article because the authors are the same neural network modeling. However, they are utilizing a single layer whereas my model is using four layers: an input layer, two hidden layers and an output layers

***Title: Single-hidden layer neural networks for forecasting intermittent demand***

In this article, the authors are proposing a method that uses a single-hidden layer feedforward neural network for forecasting intermittent demand. This is based on the back-propagation gradient-descent, perceptron algorithm described in the algorithm overview section. However, due to some of the drawbacks of back-propagation gradient descent algorithms including, slow convergence, setting learning parameters, etc. They are relying on a “faster learning algorithm”, first introduced by *Huang et al. (2006),* and later reviewed by the same authors *Huang et al. (2015).* Their proposal aims to “helping the network to learn the temporal behavior of the time series in terms of zero/non-zero demand” *(Lollia et al. 2017).* They compared the accuracy of their forecasting model with other Neural Networks that dealt with intermittent demands *(Guttierez et al., 2008; Mukhopadhyay et al. 2012; Croston, 1972; Syntetos and Boylan, 2005)*

As conclusion, the authors argue that “this comparison was then enriched by adopting two different accuracy metrics on different time horizons. Such a detailed comparison aims at bridging the gap between theory and practice of ANNs in the field of intermittent demand. *(Lollia et al. 2017)*. In fact, the potential for implementation of ANNs in real environments can only increase by providing useful guidelines about their design and training for practitioners. Finally, a statistical analysis of the networks’ performance was conducted, for robust validation of the results.” *(Lollia et al. 2017)*

**Novel Contribution**

This work is an effort to contribute to the research and the exploration of Neural Networks. The contribution is to help understand Neural Networks better and show how they can be successfully applied to stock market data analysis. The proposal is inspired by the previous excellent work that has been done by data scientists and other professionals using Neural Networks. My idea is to use R programming packages LM and NEURAL NETWORK to perform a comparative analysis of linear regression and neural network time series analysis to build a model to predict and forecast stock market volatility. The ultimately goal is to build a high performant stock market analysis compact model that can be used as an investment tool.

**Data Source and Content**

The original full data set that I am using was provided by Professor Robert Finn last semester during our DS-640 course. It is a stock market data set. It contains data about Income Statement, Cash Flow Statement, Balance Sheet, and Metrics and Ratios.

I am only using a subset of the data corresponding to the ARQ (As Reported Quarterly) listings, meaning where Dimension = ARQ. The Rows are the quarterly calendar date Time Series. The Columns are twenty chosen indicators along with the calculated Returns and Log of returns.

My list of twenty chosen indicators or factors is the following:

1. CURRENTRATIO Current Ratio
2. DE Debt to Equity Ratio
3. DIVYIELD Dividend Yield
4. FCFPS Free Cash Flow per Share
5. GROSSMARGIN Gross Margin
6. PAYOUTRATIO Payout Ratio
7. SPS Sales per Share
8. NETMARGIN Profit Margin
9. BVPS Book Value per Share
10. EVEBITDA Enterprise Value over EBITDA
11. PE1 Price to Earnings Ratio
12. PS1 Price to Sales Ratio
13. TBVPS Tangible Asset Book Value per Share
14. PRICE Share Price (Adjusted Close)
15. PS Price Sales Damodaran Method
16. EBITDAMARGIN EBITDA Margin
17. PB Price to Book Value
18. REVENUE Revenues
19. EPS Earnings per Basic Share
20. PE Price Earnings Damodaran Method

**My Method**

For implementation, R programming packages LM and NEURAL NETWORK are used. The NEURAL NETWORK package is based on the back-propagation gradient descent algorithm. I am utilizing a double-hidden layer neural network for training.

I first used an R for loop to compute the returns from price using the formula: Returns = P(i)/P(i-1). I added it to the data set as separate column. After selecting the twenty indicators along with two columns: ticker and calendar date, I then used another R for loop to compute the log of returns and added it to the data set as a separate column.

I finally normalized all the factors using the following formula:

Normalized factor = (factor – mean (factor))/sd (factor), where mean is the mean of the column and sd is the standard deviation of the column.

I will write R code to generate time series for the 20 betas derived from my linear regression of log-returns on your 20 chosen factors over the first three quarters of the dates for which I have data. I will then use these time series to forecast the 20 betas for the first date not included in the model which will give me an expected return for each stock in my model over the next time period. This expected return will then be employed to rank my stocks and split them into 5 groups with group 1 having the highest expected return and group 5 the lowest expected returns. The average actual returns for each group will then be calculated for the first date not included in the model. I will then continue this process on a rolling basis for the rest of the dates for which I have data. I will perform the same task using a neural network instead of a linear model.

**Quantitative Results**

1. For each of the dates not included in the model, meaning the fourth quarter of each the years 2011 and 2012, I will predict the return for each stock with the neural network model, and use that to rank my stocks in a decreasing order before splitting them into 5 groups. I will then calculate the average returns for each group, and discuss the difference in returns for each group.
2. For the fourth quarter of 2011.

|  |  |  |
| --- | --- | --- |
| Groups | Average returns for 2011-12-31 | Predicted returns for 2011-12-31 |
| Group 1 | 0.11920804902 | 0.10329895143 |
| Group 2 | 0.11864927415 | 0.08354768423 |
| Group 3 | 0.10646070662 | 0.07487498111 |
| Group 4 | 0.05638041033 | 0.05834756905 |
| Group 5 | -0.02173637167 | -0.00277221546 |

**Table 1.1: Compared returns for 2011-12-31**

1. For the fourth quarter of 2012.

|  |  |  |
| --- | --- | --- |
| Groups | Average returns for 2012-12-31 | Predicted returns for 2012-12-31 |
| Group 1 | 0.114665701169 | 0.11928866714 |
| Group 2 | 0.115336406124 | 0.10229519698 |
| Group 3 | 0.127731203722 | 0.09334380542 |
| Group 4 | 0.093139804160 | 0.08427562973 |
| Group 5 | -0.008158161041 | 0.03373893868 |

**Table 1.2: Compared returns for 2012-12-31**

1. For each of the dates not included in the model, for the consumer services sector, I will predict the return for each stock in the neural network model, and use that to rank my stocks in a decreasing order before splitting them into 5 groups. I will then calculate the average returns for each group, and discuss the difference in returns for each group. My dates that are omitted from the model are the fourth quarter years 2011 and 2012.
2. For the fourth quarter of 2011.

|  |  |  |
| --- | --- | --- |
| Groups | Average returns for the consumer services sector for 2011-12-31 | Predicted returns for the consumer services sector for 2011-12-31 |
| Group 1 | 0.07440042386 | 0.19869675685 |
| Group 2 | 0.10258814887 | 0.10096677348 |
| Group 3 | 0.11422949794 | 0.07340414277 |
| Group 4 | 0.14370097771 | 0.03797191082 |
| Group 5 | 0.11073043774 | -0.08251064998 |

**Table 2.1: Compared returns for the consumer services sector for 2011-12-31**

1. For the fourth quarter of 2012.

|  |  |  |
| --- | --- | --- |
| Groups | Average returns for the consumer services sector for 2012-12-31 | Predicted returns for the consumer services sector for 2012-12-31 |
| Group 1 | 0.07292894974 | 0.18933516259 |
| Group 2 | 0.11057367223 | 0.12152389766 |
| Group 3 | 0.10853561096 | 0.10101327971 |
| Group 4 | 0.10615710824 | 0.07756437083 |
| Group 5 | 0.10503236729 | -0.04601719844 |

**Table 2.2: Compared returns for the consumer services sector for 2012-12-31**

**Discussion: Comparison with My Competitor**

The authors argue that “this comparison was then enriched by adopting two different accuracy metrics on different time horizons. Such a detailed comparison aims at bridging the gap between theory and practice of ANNs in the field of intermittent demand. In fact, the potential for implementation of ANNs in real environments can only increase by providing useful guidelines about their design and training for practitioners. *(Lollia et al. 2017)*.

**Discussion**

1. **Performance of the model**
2. For the fourth quarter of 2011, a comparison between the average log of returns and the predicted returns using neural network shows that they are lower and pretty close for group 1 where the average return is 0.12 and the predicted return is 0.10. For group 2, predictions are little lower, with an average return of 0.12 and a predicted return of 0.08. We see the same trend for group 3 with an average return of 0.11 and a predicted return of 0.07. Predictions are close and a little higher for group 4 with an average return of 0.056 and a predicted return of 0.058. Predictions for group 5 are higher with an average return of -0.0217 and a predicted return of -0.0028. Returns for group 5 remain in negative territory for this quarter.
3. For the fourth quarter of 2012, a comparison between the average log of returns and the predicted returns using neural network shows that overall they are pretty close for all the groups. Predictions are higher for group 1 with an average return of 0.11 and a predicted return of 0.12. For group 2, predictions are little lower, with an average return of 0.12 and a predicted return of 0.10. We see the same trend for group 3 with an average return of 0.13 and a predicted return of 0.09. Predictions are lower but very close for group 4 with an average return of 0.09 and a predicted return of 0.08. Predictions are higher for group 5, with an average return of -0.008 and a predicted return of 0.034.
4. **Applying the model to the Consumer Services Sector**
5. For the fourth quarter of 2011, a comparison between the average log of returns and the predicted returns using neural network shows that overall they are pretty close for all the groups. Predictions are higher for group 1 with an average return of 0.07 and a predicted return of 0.20. For group 2, predictions are little lower, with an average return of 0.1026 and a predicted return of 0.1010. Predictions are lower for group 3 with an average return of 0.11 and a predicted return of 0.07. Predictions are lower for group 4 with an average return of 0.14 and a predicted return of 0.04. We see a similar trend for group 5, lower and close, with an average return of 0.11 and a predicted return of -0.08. Predictions for group 5 reached negative territory for this quarter.
6. For the fourth quarter of 2012, a comparison between the average log of returns and the predicted returns using neural network shows that overall they are pretty close for all the groups. Predictions are little higher for group 1 with an average return of 0.07 and a predicted return of 0.19. For group 2, predictions are little higher, with an average return of 0.11 and a predicted return of 0.12. Predictions are lower but very close for group 3 with an average return of 0.11 and a predicted return of 0.10. We have similar trend for group 4 with an average return of 0.11 and a predicted return of 0.08. Predictions for group 5 are lower, with an average return of 0.11 and a predicted return of -0.05. Predictions for group 5 went from positive to negative for this quarter.

**Performance on Big Data: Time Measurements**

For loop to estimate returns based on price: *47.83 sec elapsed*

Running/training the neural network model took most of the time: *201.605 sec elapsed*

**Conclusion**

Overall, the double-hidden layer Neural Network model has performed better than Linear Regression model. Predicted returns are closer to the actual average returns using neural network. The only limitation is that it takes time to train the neural network, which leads to a slow convergence. If we increase the threshold during the training of the neural network, it will converge faster but the precision of the predictions will be affected. Many more trainings of the neural network should be conducted in order for the model to improve.

**Questions:**

* What statistical method was used?
* What programming tool was used?
* What scientific contribution was made?
* What idea could be useful for your project?