**DS-670-Capstone: Big Data & Business Analytics**

**Assignment 15: Final Tele Prompter Script**

**May 6th 2017, Madhumita D**

**Title: “Market Efficiency of U.S. Stock Market: A Test for Semi-Strong Form Efficiency”**

**Contribution of Competitor’s Article:**

“**Stock return predictability and the adaptive markets hypothesis: Evidence from century-long US data.**” is the competitor article chosen to compare the methods and analysis done in our paper. Stock return predictability measures from regression analysis were used to validate the adaptive market hypothesis which states that returns are subject to market conditions. The article discusses about various areas which help explore the return predictability in AMH environment. These have been stated below:

* Time-varying stock return predictability
* Returns are driven by changing market conditions
* Explores the implications of AMH:

Degree of market efficiency fluctuates over time.

Degree of market efficiency depends on market conditions.

* Predicted returns are consistent with the implication of adaptive market hypothesis (AMH).
* Stock returns are highly related to market volatility and economic fundamentals.
* Assumes that market is *weak-form of EMH* and states that asset returns cannot be predicted accurately using the past prices.

In this article they have used regression modelling to understand the return predictability. Regression was done on the DJIA index values and also the measures of return predictability. The article uses the following statistical tests to compute measures of return predictability:

* Automatic variance ratio test (testing weak-form of efficiency of financial markets)
* Automatic portmanteau test (testing the auto-correlation of returns which are subject to unknown forms of heteroskedasticity)
* Generalized spectral test (both linear and non-linear dependence are tested)

**Description of Your Contribution:**

The purpose of this analysis is to test efficient market hypothesis in a semi-strong efficient market using the closing prices of U.S. Stock market data. These tests are related based on different sectors of an industry to draw conclusions on the consistency of the predicted stock prices based on each sector. Stock return prediction is done using a hybrid model of neural networks and time series modelling.

Efficient market hypothesis (EMH) states that it is not possible to consistently outperform the market—which is usually based on the composite decisions of millions of investors with similar goals and equal access to the same publically available information. In a more general sense EMH explained by Eugene Fama states that the stock prices follow a random walk theory and cannot be predicted based on their past behaviour. Here “efficiency” in hypothesis means that the market can quickly digest any new information available on the company, industry, economy, or value of the company and can impound this information accurately into securities prices. Hence in efficient capital markets we can expect people to earn fair returns for the risks taken, no more and no less.

Our analysis is to test the semi strong market efficiency of pricing data on daily and quarterly observations of U.S. stock market. The study also tries to relate the same tests based on the sectors of industry to draw conclusions based on the consistency of the predicted stock prices based on each sector. We have considered using neural networks model to test the semi strong form of efficient markets by predicting the returns using the pricing information on stocks. The data is categorized mainly based on the quarters and daily stock information. Stock prices and information related to capital markets is time dependent data and forecasting the future prices is majorly dependent on historical prices and all publically available information on the stocks. Pricing information can be considered as a time series which is defined as a sequence of values measured over a period of time (discrete or continuous). Financial information of a company is released publically every year in the form of quarterly and annual statements for a fiscal year. These statements include costs, revenues, expenses, price, earnings and many other factors. Predicting stock prices based on many related factors gives a better result. So we consider a multivariate time series which consists of variables whose values change over time. Hence a multivariate time series analysis is used to study the behaviour of time dependent pricing data and forecast values based on the history of variation of the data.

**Data Source and Content:**

Data accumulation and parsing is a recursive process and when we have to analyze financial data then collection of data is an on-going process. The dataset considered is a combination of both profit and loss statement and balance sheet. P&l statement (mostly known as income statement) has indicators which help us in identifying the performance and profitability of a company whereas Balance Sheet has indicators which help us in identifying the financial status of each company. Thus using these indicators from both income statement and balance sheets we try to forecast the stocks with better expected returns. The analysis is done based on the various frequencies (daily, monthly and quarterly) available in the data set. We have considered daily closing prices, volume, fundamental factors and technical factors (RSI – Relative Strength Index) for our analysis.

**QUATERLY DATA ACCUMULATION**

* Fundamental factors are captured from financial statements which are released at the end of every quarter for a fiscal year.
* Financial statements are a combination of both profit and loss statement and balance sheet.

**DAILY DATA ACCUMULATION**

* Market capital is the company’s market value of its outstanding shares. Based on the company’s market capitalization its stocks are classified into small cap, mid cap and large cap stocks.
* Our analysis uses the stock market data of the U.S. market from the year 2011.
* Unlike the article we are using the daily closing prices of an exchange traded fund (ETF) of small cap stocks and large cap stocks (SP500).
* RSI is computed based on these prices for each stock.

**Your Method:**

Neural Networks are valued as one of the foremost techniques in performing nonlinear forecasting tasks and complicated pattern recognition across many applications and domains. Some such interesting applications of neural networks are in decoding protein sequences and deterministic chaos. With these successful applications in varied domains, it is interesting to know whether applying this technique to economic time series will yield in extracting the non-linear regularities of the economic market. Decoding the regularities in asset price movement like day to day or month to month fluctuations of stock prices (which were undetected previously) will be the key to earn benefits from investing.

The prediction of expected returns using the stock price information of U.S. market is made up of many neural networks that learn the relationship between technical and economical factors. The aim is to forecast the daily and quarterly returns of the all the stocks. Figure A depicts the basic architecture of the prediction system. It basically converts all the factors considered into a space pattern which form the input to the neural network.

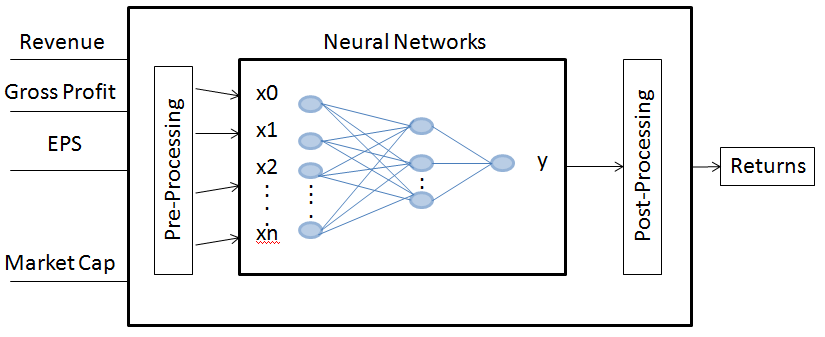


Figure A: Basic architecture of predicted system

Time series analysis is done using ARMA models. The ARMA model or autoregressive moving average model is generally applied to time series data. Most commonly it is used to compute and estimate the Value at Risk (VAR), an estimation model used in simulation forecasting of asset prices. The ARMA model can be used to forecast the future values for a given time series data. Forecasting future stock prices is based on the past prices and volume information which can be converted to a time series data.

The following steps have been taken in order to implement the method:

* Input to a neural network should be normalized first.
* Initially frequency of the model chosen is quarterly.
* For each quarter from 2011 to 2014 stock data is considered as train data and from 2015 as test data for modelling the neural networks
* Neural Network Model is run for each sector in the train data.
* Time series analysis is performed on the weights from each neural network model.
* Weights for the future quarters are predicted using time series model.
* Using the predicted weights, compute the expected returns of all the stocks.
* Stocks are ranked according to expected returns and divided into 5 groups.

The flexibility and efficiency of the neural networks gives it the ability to deal with robust and uncertain data. This means that to forecast the prices or returns of an asset, neural networks in combination with time series analysis can be efficiently used in stock market analysis. Many previous studies have shown that neural networks outperform linear models, statistical models and classical forecasting. Combining both features of neural networks and ARMA time series model, building a network for prediction of returns is much more complicated. In order to get a robust model all the crucial factors must be considered while designing a prediction model.

|  |  |
| --- | --- |
| Notation | Description |
| Pt | Stock Price |
| Rt | Return = (Pt /Pt-1) – 1 |

Table A: Explanatory variables

The hybrid approach of using neural networks with time series analysis addresses the fundamental problems by predicting a noisy time series. Predicting stock prices in an economic system is subject to change vigorously. For such forecasting the prediction rules continuously change and hence learning and prediction must adhere to these changing rules. A prediction system known as moving simulation was introduced for this kind of changes. In moving simulation method the prediction is carried out by simulation while moving the objective learning and prediction time intervals.

**Quantitative Results:**

The data is analyzed after the neural network model for each quarter was established. The expected returns were calculated based on these models for the test data. The neural network model for every quarter had 4 layers – input layer consisting of the technical, fundamental and macro-economic factors, two hidden layers with six and four hidden units respectively, and an output layer which results in the expected returns.

After the stock returns for test data has been computed, the stocks for every quarter are ranked based on these returns. These ranked stocks are divided into five quantiles – highest to lowest. The average of each quantile defines the average returns of stocks in that bucket. In the similar way, returns, ranking and quantiles have been computed for all the future five quarters. Based on every quarter’s average returns of each quantile we will be able to identify the trends in each quantile over a period. This will help us in identifying whether there was a positive trend or negative trend followed in a particular quarter. The difference between top quantile and the bottom quantile helps us in identifying the dispersion in returns in a quarter. Also, we can identify the expected returns of top quantiles or bottom quantile in each quarter which helps us in making investment decisions (buy or sell). To demonstrate the above theory in a particular sector we have used the proposed model in Finance sector for quarters – Q4 2014, Q1 2015, Q2 2015, Q3 2015 and Q4 2015.

The results from the quarterly stock analysis (Figure B) have been explained in detail. Based on the overall forecasted values we can say that the expected return of financial stocks have shown a positive trend in all the quarters with Group 1 stocks out performing Group 5 stocks. This can be observed in both predicted values and actual values.

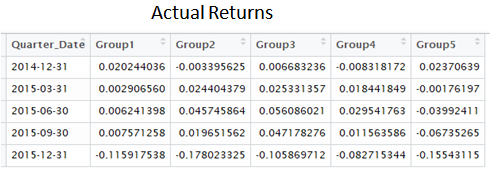
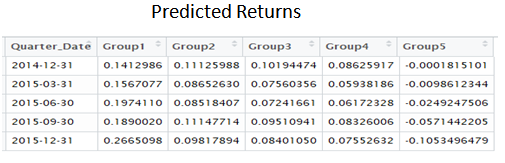


Figure B: Quarterly Average Predicted and Actual Returns for each quantile.

The spread between the performance of group 1 and group 5 (Figure C) has increased for every quarter. On the whole the averages depict a huge volatility in stock price movements. The return predictability is dependent on market conditions and reflects all the publically available information which is evidence that the U.S. stock markets are semi-strong form efficient.

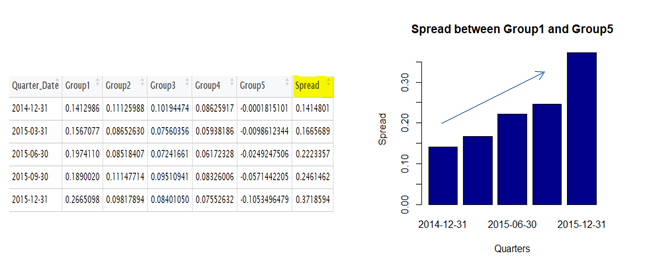


Figure C: Spread between Group1 and Group5 for each quarter. Group1 – well performed stocks and Group 5 – poorly performed stocks.

**Discussion: Comparison with Your Competitor:**

Traditional methods like regression analysis for prediction have their limitations in the applications with data sets which show nonlinearities. Many forecasting techniques are capable only of selecting general trends like positive and negative trends, and show complexity in modelling cycles which by no means are repetitive in period, amplitude, or shape. Even though such inadequacies are present, modelling techniques such as multivariate linear regression is used routinely in the Capital Markets and also have proved to be a very useful tool.

In our paper we have showed that using a simple neural network employing publically available data we are able to successfully forecast a group of stocks which will out-perform the market as well as a group of stocks which will under-perform the market, directly contradicting the semi-strong form of the EMH. With their smooth interpolation properties neural networks allow models to fit better to the data and generalize significantly better.

We have used a hybrid approach of neural networks and time series analysis to calculate the return predictability. Based on the overall forecasted values we can say that the expected return of stocks have shown a positive trend in all the quarters. The spread between the performance of group 1 and group 5 has increased for every quarter. On the whole the averages depict a huge volatility in stock price movements.

We have assumed that the price for each stock is highly influenced by the fundamental, technical and macro-economic factors while the regression analysis (used in competitor’s article) was done under the assumption that the market follows a weak-form of efficient market hypothesis (EMH). This assumption was made to test for evidence of AMH (adaptive market hypothesis) in return predictability. The hybrid approach followed in this study helped to test for semi-strong form of efficient market hypothesis (EMH) based on return predictability. The returns were computed from the pricing information for each stock and predicted for the future time using the neural networks and time series analysis.

Regression analysis in (competitor’s article) uses Dow Jones Industrial Average (DJIA) as a technical indicator. The index is a price-weighted average of 30 blue-chip stocks, accounting for 25–30% of the total value of U.S. stocks. Index created for 30 blue chip stocks (large cap stocks) for century-long U.S. stock market data from 1900 to 2009. We have used daily closing prices of both small and large cap stocks captured from exchange traded fund (ETF) for the duration of 2011 to 2016 for U.S. stock market. In our model Relative Strength Index (RSI) is used as a technical indicator. Relative Strength Index compares the magnitude of recent gains and losses over a period of time based on the change of price movements.

Neural Networks and Factor Models were implemented to forecast the returns of a stock in the market. We use a multivariate time series to study the behaviour of stocks. To weigh the use of neural networks we compare the results and metrics to Regression Analysis, which was conducted on the measures of return predictability of stocks. Many different metrics were used to compare and compute the results. In Neural Networks sequential testing of market efficiency plays a major role in computing t the returns accurately based on all the historical and present data available while sequential testing of market efficiency was not taken into consideration in linear regression.

Our analysis addresses that the factors chosen drive the performance of the model but also the Investability of our model with real-time data. This entails simulation of a portfolio controlling for risk and liquidity of the assets traded. We showed that predicted returns reflect the publically available information and are dependent on market conditions which prove that markets are semi-strong form efficient.

**Performance on Big Data: Time Measurements:**

The table below shows the performance (as related to time) of each operation used in our analysis. On the whole it took about 55 minutes to complete this analysis.

|  |  |
| --- | --- |
| **Operation** | **Time Measurement** |
| Data Load | 240 seconds |
| Merging daily files with quarterly data | 60 seconds |
| Calculating RSI values for each file | 60 seconds |
| Normalizing the indicators values for each file | 3 seconds |
| Neural Networking modelling | 45 minutes (each model takes about 5 minutes) |
| Time Series Analysis on weights | 5 seconds |
| Predicting Returns for future quarters (using the neural network equation) | 120 seconds |
| Ranking the datasets (using the expected returns) | 5 seconds |
| Grouping the stocks into quantiles | 5 seconds |
| Average of stocks in each quantile | 5 seconds |

**Conclusion:**

The hybrid approach combining Neural Networks and Time-Series used in this study was able to categorize between the stocks that performed well and stocks which performed poorly in the market. This approach also outperformed the multiple linear regression analysis approach overall. Investors who use this modelling technology should be able to improve their analysis choosing the best investment alternatives based on better predictions. However, further research is required in using this hybrid approach in stock prediction analysis. To further validate the proposed technique in investment analysis, we need to perform qualitative and quantitative tests on our data, considering different time periods and different industries or sectors. With the many advantages, investors and researchers should seriously consider the application of neural network and time-series techniques to investment analysis.