Short-term prediction of exchange traded funds (ETFs) using logistic regression generated client risk profiles

Jerry K. Bilbrey, Jr.
Clemson University

Neil F. Riley
Francis Marion University

Caitlin L. Sams
Anderson University

ABSTRACT

Efficient markets are a major tenet of investment theory. Efficient markets fully reflect all information into the price of a given asset such as stocks, bonds or exchange traded funds (ETFs). This suggests that there are no abnormal profits that can be made based on known public data. This paper presents an approach that goes against the efficient market theory by presenting a method that utilizes price and volume data to predict buy and subsequent sell signals for a list of ETFs. This approach utilizes both linear and logistic regression that develops a method for generating these buy signals. Sell signals are automatically created on the risk profile generated for each ETF by the model. As detailed, the regression based model shows great promise for developing strategies using individual risk based profiles.

Keywords: Efficient market hypothesis, logistic, regression, Exchange Traded Funds

INTRODUCTION

The topic of stock market returns and the ability to get excess returns based on publicly available data has been around for an extended period of time. In fact, there have been many studies that investigated the predictability of stock market returns using linear models on publicly available data (Schwert, 1990) and (Balvers, Cosimano, and McDonald, 1990). In fact, Jordan, Vivian, and Wohar, (2012) provide a rather exhaustive literature review concerning linear models for the interested reader. However, it is widely known that stock markets often display non-linear behavior. Therefore, it is necessary to find methods that can overcome the limitations of linear models with respect to financial markets.

One choice for overcoming linear model limitations is the neural network. Vellido, Lisboa, and Vaughan (1999) state that financial forecasting has been done extensively using multi-layer feed-forward neural networks because of their inherent ability to classify and predict a given dependent variable. Hill, O'Conner, and Remus (1996) find that neural networks have the ability to overcome problems associated with linear based methods for financial market predictions. These are further explored in the literature review below.

As much research as has been done in stock market predictions, there is a paucity of work in the area of applying some nonlinear statistical models such as logistic regression. The contribution of this work is to apply logistic regression to Exchange Traded Funds which often mimic overall sectors of the market similar to mutual funds. A relatively new asset class called ETFs, is chosen because it is not as established nor has it been explored as much as many other asset classes. It is found through experimentation that risk profiles can be generated that can allow a user to choose the amount of desired risk to correspond with a given time horizon. In the rest of this paper, the hypothesis that there is predictive information in price and volume data for ETFs is explored further.

LITERATURE REVIEW

When trying to determine how to predict the stock market, researchers have found many different ways of accomplishing the task through many different types of models. Some use statistics, neural network models, basic experiments and some compare and contrast models. There are many models that have been found to be beneficial which are outlined below. First, the basic foundation of financial markets is described. Then, the current state of research based on this information is explored by looking at some nonlinear approaches.

The efficient market hypothesis claims that prices fully reflect all information. This theory of investment informational efficiency dates back to the dissertation of Louis Bachelier (1900). The early works of Fama (1965) and Samuelson (1965) provided support for the random walk hypothesis and the efficient market hypothesis. See Dimson and Mussavian (1998) for an excellent history on the development of the efficient market hypothesis.

The efficient market hypothesis is examined via various tests of market behavior and efficiency. The weak form test of the efficient market hypothesis posits that prices fully reflect all information implicit in historical price and volume information. The semi-strong form test of the hypothesis asserts that all publicly available information is reflected in market prices. The strong form test of the hypothesis asserts that information known to any market participant is fully reflected in market prices.

As computer technology caught up with investment theory, tests of market efficiency began to concentrate on the reflective efficiency of the market. That is, how quickly (efficiently) do market participants reflect new information into the price of a security? Technical analysts search for shifts in price and volume that can lead to trading rules that will signal buys and sells before all profit is lost. These technical trading rules are viewed to be most impacted by studies that continue to show support for the random walk hypothesis and the weak form test of the efficient market hypothesis. Fundamental analysts believe that searching out information about a company and applying rigorous analysis can lead one to superior returns. This type analysis does not hold up well given the results of the semi-strong form tests of the efficient market hypothesis. Most semi-strong form tests show that acting on publicly traded information does not lead to superior risk adjusted returns. The strong form tests of the efficient market hypothesis shows that some market participants do have information that may be useful. Corporate insiders may have information that allows them to legally trade the stock of their company and earn superior risk adjusted returns. Some corporate insiders' trade on information that earns superior returns, but the information is considered illegal insider information and the profits are illegal. Does the profit have to be legal to disprove the hypothesis?

Much of the criticism directed toward the supporters of the Efficient Market Hypothesis comes from the professional investment community. Accurately noting that many academic studies are based on statistical tests that are linear, the argument has been that the factors driving risk, return, and ultimately pricing of securities is much more complex. Lendassse et al. (2000) provide linear and non-linear models to show the ability of a neural network to provide results which "...seem sufficiently strong to question the random walk hypothesis."

There have been many attempts at quantifying neural network approaches for investments with traditional investment methods. Gençay (1998) presents a nonparametric model to maximize profits from an investment strategy. It is shown that when compared with a simple buy-and-hold strategy the technical strategies provide significant profits. Also, it is reported that the sign predictions of the nonparametric models are statistically significant. Enke et al. (2005) present an information gain technique to evaluate the predictive relationships of some financial and economic variables. They also examine the differences between level-estimation neural networks and probabilistic neural networks and conclude that models, such as a probabilistic neural network, provide higher risk-adjusted profits than a simple buy-and-hold strategy in a financial market. Thawornwong et al. (2003) investigate using technical indicators as input to a neural network. Their results indicate that the proportion of correct predictions and the profitability of stock trading are higher than traditional benchmarks.

Similarly, there have been attempts to compare neural network methods with other types of quantitative methods. Jasic and Wood (2004) report that neural networks provide significant information gain over a benchmark linear autoregressive model. Also, they show that buy and sell signals provided by neural networks are significantly different from unconditional one-day mean returns. Olson and Mossman (2003) compare neural network forecasts against the forecasts of ordinary least squares and logistic regression. They report that neural networks have superior performance over the other two methods but do not report the effectiveness of short term predictions. Bilbrey et al. (2007) report similar results using a probabilistic neural network. They indicate that the neural network performs better than the large cap index but not as well as the small-cap index over their time horizon.

Leung et al. (2000) evaluate the efficacy of several classification techniques in financial markets which include probabilistic neural networks, adaptive exponential smoothing, logit and

discriminant analysis. Their results indicate that probabilistic neural networks perform better, almost exclusively, over the other methods tested. In addition, they note that using two thresholds for a trading strategy works better than a single threshold. For instance, when using a probabilistic neural network, a single threshold purchase strategy could be used when the network outputs a value 0.5 or higher. However, a multiple threshold strategy allows for purchasing securities, shorting the market, or even doing nothing.

Jordan, Vivian and Wohar (2012) investigate the question of linear regression based models to enhance allocation decisions in real-time as applied to international data. This approach generated substantial economic value compared to an approach using a purely statistical approach based on the mean squared error. They found this to be true "across several methodologies, including individual forecasts, bagging methods, and combination methods."

Poon and Taylor (1992) find it useful to use the UK's financials to estimate the stock market vs. volatility. The research included their results which stated that they found evidence for a negative relationship but only when volatility expectations were represented by standard deviations. Standard deviations are important to know because they help finding a variation close to the average (mean). Ding et al. (1993) use standard deviations to predict the stock market. They study the "long memory" property of the stock market. They state that the study of "long memory" is enough to say that it is the strongest when the standard deviation is close to 1. They use the heteroskedasticity equation to estimate their results from the data. These results are helpful when deciding how to predict the stock market because it makes for finding standard deviations important.

Maasoumi and Racine (2002) study the use of the entropy to measure the stock market. They use a model assuming that $100 is invested in stocks or bonds with either low or high cost of transactions. Low transaction costs are .5% on trading stocks and .1% on trading bonds, while high cost transactions are 1.0% on trading stocks and .1% on trading bonds. After comparing linear to nonlinear returns the authors conclude that a use of low and high transaction cost were beneficial. Another way to use a linear model is by predicting quarterly stock market excess returns. Hiemstra (1996) stated that quartely stock markets returns are somewhat predictable. The ending results confirm from other studies that a fundamental model using a limited number of inputs has predictive power.

Rapach and Wohar (2006) use S&P 500 and CRSP equal-weighted real stock returns based on 8 financial variables to predict regression models. The authors use Andrews SUPF statistics and the BAI subsample procedure in conjunction with the Hansen heteroskedastic fixed-regressor bootstrap to test for structure stability. In conclusion, they found strong evidence of structural breaks in 5/8 bivariate predictive regression models of S&P 500 returns and some evidence of structural breaks in the other 3 models. Campell (2007) also used the S&P index. Campell begins by conducting an out-of-sample forecasting exercise inspired by Goyal and Welch (2008) with modifications that reveal the effectiveness of theoretically motivated restrictions. He uses monthly data to predict simple monthly or annual stock returns on the S&P 500 Index.

Todd and Correa (2007) present 2 experiments using the Gaussian process. They keep track of important stock information such as: price and whether the price was increasing or decreasing by useing a + 1 labeling to produce a simple metric trend. (1 when increasing and -1 when decreasing). Next, he looked to see if the value of the stocks were close to the trading day so they could predict the price of the stock for the following days. To help his experiment he used 100 iteration optimization of the marginal liklehood with rach different kernel. M. H.

Pesaran, author of Market Efficiency and Stock Market Predictability (2003), stated he found by using the return regression equation was statistically significant.Evidence of stock market predictability can be done by using intrest rates, dividend yields and a variety of macroeconomic variables exhibiting clear business cycle variations.

It has been shown that there is evidence to argue against the efficient market hypothesis. There is also substantial and growing evidence that financial markets display complex behavior that causes linear models to be less suitable for market prediction than neural network models or other types of nonlinear models. Overall, these authors have been successful in their research. Whether the use of standard deviations, linear and rarely nonlinear regression were used, all authors came up with similar results from their experiments. It does appear that there is reasonable predictability for many models with stock market applications. Although this conclusion looks rather sound, there has been a minimal amount of research conerning logistic regression and its application to the stock market. This paper continues the pursuit of finding predictability in the stock market while using a relatively unexplored tool – logistic regression.

METHODOLOGY

Survivorship

The method for choosing the ETFs under consideration went through a method of survivorship. The survivorship period under consideration was 2007-2012. First, the MLDownloader software (Trading-tools.com, 2012) was used to download data for all ETFs. Then, each of the ETFs was surveyed to determine which ones had data that covered the five year period under consideration. As ETFs are one of the more recently created security classes, using the period was somewhat limiting. However, there was a significant number of ETFs that were available using this constraint. It was believed that a model would be more reliable if there were a minimum of two years of input data. Out of approximately 500 ETFs, 128 had enough available data to survive the process. Out of these 128 ETFs, about 20 of them indicated that there was significant predictability by the modeling process. This was determined through the ranges of predictability by the generated risk profiles.

Modeling Process

Logistic regression is a non-linear statistical tool that allows the regression modeling of an output variable in terms of a probabilities, logged odds and odds with interpretations of each effect that has both advantages and disadvantages (Pampel, p. 18). For this study, the chosen output of the logistic regression is given in terms of probability. This is chosen because it is a well understood terminology throughout industry whereas odds are somewhat understood and logged odds are not well understood at all. This allows the model to be utilized by a wide range of users that do not need specialized training to understand and convert the output. Also, since a risk profile is created, it is very similar to the "multiple threshold strategy" described by Leung et al. (2000). This allows the user to select a personal threshold strategy based on the predicted probability of success.

Logisitc regression is quite similar to use compared with linear multiple regression but it gives the added dimension of nonlinear behavior. For example, equation 1 generated for a two variable linear multiple regression would look like the following.

$$Y = \alpha + b_1X_1 + b_2X_2 \qquad \text{equation (1)}$$

where X_1 and X_2 are independent variables
Y is the predicted or explained dependent variable
α is the *Y* intercept
b_1 is the the net change in *Y* for each unit change in X_1 holding X_2 constant

This is the basis equation for a regression involving two variables. Extending this equation to include an interaction would look like equation 2 below.

$$Y = \alpha + b_1X_1 + b_2X_2 + b_3X_1X_2 \qquad \text{equation (2)}$$

This is the basis format for modeling the relationships where X_1 is Price, X_2 is Volume, and X_1X_2 is the interaction term. This equation is very similar to the final logistic regression equation but it is actually extended to a total of 15 variables with three parts each as shown in equation 2. First, an explanation of logistic regression is appropriate. Logistic regression has the ability to create a probability of success (user defined success) based on empirical data of successes and failures. The output follows an S-curve pattern that gives less dramatic changes toward the extremes of 0 or 1 because of the curve of the output. The logistic regression curve becomes asymptotic to the horizontal axes as it becomes closer to 0 or 1 respectively. The basic equation for the probability of output of a logistic regression equation for two independent variables with one interaction term would look like equation 3.

$$P = \frac{e^{(\alpha + b_1X_1 + b_2X_2 + b_3X_1X_2)}}{1 + e^{(\alpha + b_1X_1 + b_2X_2 + b_3X_1X_2)}} \qquad \text{equation (3)}$$

Custom software was created to test the given hypothesis that there is predictive information in price and volume data for ETFs. The C++ language was used to build the model and simulate the results for each ETF. Minitab Statistical Software version 14.0 (Minitab Inc., 2005) was used to generate the logistic regression equations that were input into the custom software model to test the hypothesis. The logistic regression coefficients are generated by the statistical software. The resulting regression equations are then input into equation 3 (within the C++ model) to generate a probability of success. Based on the risk profiles (described below) the user can run the model to determine which ETFs satisfy the current conditions for success based on the empirical data.

There were several issues during the modeling process that are worth mentioning for future researchers. As previously mentioned, the data had to pass a survivorship criterion. This rejected all ETFs that did not have five years of data. Two ways to increase the number of ETFs for future studies would be to reduce the survivorship period or to simply rerun the analysis in a few years because more ETFs would be five or more years old at that time. The five years of data were first imported using MLdownloader into Microsoft Excel files. These files turned out to be quite large for the manual process of copying and pasting into the Minitab software. As such, the copying and pasting process took several minutes per file to get paste the data into Minitab to begin the process of creating the logistic regression equation for each ETF. Multiplying this process for each of the 128 "survived" ETFs made this process extremely

cumbersome and slow. However, it was a needed part of the process to create the logistic regression equation for each ETF.

CONCLUSIONS

The final product is a family of risk profiles generated through the empirical data that can be utilized in the decision making process. The modeling process required running the model for each input ETF to determine the cutoff value for each security. Cutoff values were chosen based on the step that generated up to 600 buy signals over the empircal data time frame.

A sample risk profile for the SPDR S&P Biotech ETF with symbol XBI is shown in Table 1. An end user can utilize this risk profile whenever the C++ model generates a buy signal for a given ETF. For instance, the cutoff value for XBI is .999 with a total number of buy signals of 536.

This value is used by the C++ model to determine if the security should be given in the output as a buy for the period. Once a buy signal is created, a user can choose their level of risk tolerance for the given ETFs and purchase with an expected return given a past probability of that return. Of course, as the rate of return goes up, the probability of success goes down. These risk profiles were done for each of the 128 ETFs in the study.

To continue on with the example, after a generated buy signal is produced for a security like XBI, an end user could purchase the security based on their personal desired risk level. For instance, the user could purchase XBI for their risk preference with an expectation this return would be achieved with a given probability. For instance, the model predicts that there is a 75.8% chance that a desired 1% gain will be achieved over the next five trading days. This percentage is based completely on the model's simulated output against the empirical data.

The model covers approximately five years of data (2007-2012). During this time frame which included a major recession the model accurately predicts certain ETFs that appear to be likely to generate superior returns over the tested 5 and 10 day holding periods. This is done by creating probabilities for success using the logistic regression equations fit by the modeling process. The results of the model are predicted ETFs that a user can purchase based on their personal risk preferences and corresponding risk profiles for the given ETFs.

REFERENCES

Bachelier, L. (1900). trans. James Boness. Theory of Speculation. in Cootner (1964), pp. 17-78.

Balvers, R., Cosimano, T. & McDonald, B. (1990). Predicting Stock Returns in an Efficient Market. *Journal of Finance*. 55(4). pp. 1177-1189.

Jerry K. Bilbrey, Jr., Neil F. Riley, and Winnie A. Riley. (2009). A Test of Market Efficiency using a Probabilistic Neural Network. *Journal of Business and Behavioral Sciences*. 20(1). spring.

Campell, J. (2007). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *The Review of Financial Studies.* 21(4). pp. 1509-1531.

Cichocki, A., Stansell, S., Leonowicz, Z., & Buck, J. (2005). Independent Variable Selection: Application of Independent Component Analysis to Forecasting a Stock Index. *Journal of Asset Management.* 6(4). pp. 248-259.

Dimson, E., & Mussavian, M. (1998). A Brief History of Market Efficiency. *European Financial Management.* 4(1). pp. 91-193.

Ding, Z., Granger, C.W. & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of empirical finance*. 1(1). pp. 83-106.

Enke, D., & Thawornwong, S. (2005). The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns. *Expert Systems with Applications.* 29(4). pp. 927-940.

Fama, E. (1965). The Behavior of Stock Market Prices. *Jounal of Business.* 38. pp. 34-105.

Gençay, R. (1998). Optimization of Technical Trading Strategies and the Profitability in Security Markets. *Economics Letters.* 59(2). pp. 249-254.

Hiemstra, Y., (1996), Linear regression versus back propagation networks to predict quarterly stock market excess returns, *Computational Economics*, 9(1), pp. 67-76.

Hill, T., O'Connor, M., & Remus, W. (1996). Neural Network Models for Time Series Forecasts. *Management Science.* 42(7). pp. 1082-1092.

Jasic, T. & Wood, D. (2004). The Profitability of Daily Stock Market Indices Trades Based on Neural Network Predictions: Case Study for the S&P 500, the DAX, the TOPIX, and the FTSE in the Period 1965-1999. *Applied Financial Economics*. 14. pp. 285-297.

Jordan, S., Vivian, A. & Wohar, M. (2012), Forecasting Asian Market Returns: Bagging or Combining?, Proceedings of *International Symposium on Forecasting*, Boston, June.

Lendasse, A., De Bodt, E., Wertz, V., & Verleysen, M. (2000). Non-linear Financial Time Series Forecasting – Application to the Bel 20 Stock Market Index. *European Journal of Economic and Social Systems*, 14(1), pp. 81-91.

Leung, M., Daouk, H. & Chen, A. (2000). Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models. *International Journal of Forecasting*. 16(2). pp. 173-190.

Maasoumi, E. & Racine, J. (2002). Entropy and predictability of stock market returns. *Journal of Econometrics.* 107(1). pp. 291-312.

Marquering, W. (2004). Seasonal Predictability of Stock Market Returns. *Review of Business and Economics*. 47(4). pp. 557-576.

Minitab Inc. (2005). Minitab Statistical Software (Version 14.20) [Computer Software]. Available from www.minitab.com.

Olson, D. & Mossman, C. (2003). Neural Network Forecasts of Canadian Stock Returns using Accounting Ratios. *International Journal of Forecasting.* 19(3). pp. 453-467.

Pampel, F. C. (2000). *Logistic Regression: A Primer*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-132. Thousand Oaks, CA: Sage.

Pesaran, M. H. (2003). Market Efficiency and Stock Market Predictability. *Mphil Subject 301 Lecture Notes.*

Poon, S. H., & Taylor, S.J. (1992). Stock returns and volatility: an empirical study of the UK stock market. *Journal of Banking & Finance.* 16(1). pp. 37-59.

Rapach, D. E., & Wohar, M.E. (2006). Structural breaks and predictive regression models of aggregate US stock returns. *Journal of Financial Econometrics.* 4(2). pp. 238-274.

Samuelson, P. (1965). Proof That Properly Anticipated Prices Fluctuate Randomly. *Industrial Management Review*, 6, pp. 41-49.

Schwert, W. (1990). Stock Returns and Real Activity: A Century of Evidence. *Journal of Finance*. 45. pp. 1237-1257.

Thawornwong, S., Enke, D. & Dagli, C. (2003). Neural Networks as a Decision Maker for Stock Trading: A Technical Analysis Approach. *International Journal of Smart Engineering System Design*. 5(4). pp. 313-325.

Todd F., & Correa, A. (2007). Gaussian Process Regression Models for Predicting Stock Trends. MIT technical report.

Trading-tools.com. (2012). MLDownloader (Version 7.1.0.9) [Computer Software]. Available from www.trading-tools.com.

Vellido, A., Lisboa, P. & Vaughan, J. (1999). Neural Networks in Business: A Survey of Application (1992-1998). *Expert Systems with Applications*. 17(1). pp. 51-70.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies.* 21(4). pp. 1455-1508.

Xu, Q. (2010). A New Algorithm to Forecast Shanghai Composite Index. *Journal of Information & Computational Science.* 7(12). pp. 2463-2467.

APPENDIX

Table 1 - Results for 5 day time span for SPDR S&P Biotech - XBI

Past Success Probability	**Risk Preference**
result: 0.909176	Percent: 0.0025
result: 0.861423	Percent: 0.005
result: 0.811798	Percent: 0.0075
result: 0.758427	Percent: 0.01
result: 0.702247	Percent: 0.0125
result: 0.645677	Percent: 0.015
result: 0.594925	Percent: 0.0175
result: 0.531015	Percent: 0.02
result: 0.491541	Percent: 0.0225