Machine Learning Techniques for Stock Prediction

Vatsal H. Shah

1. Introduction

1.1 An informal Introduction to Stock Market Prediction

Recently, a lot of interesting work has been done in the area of applying Machine Learning Algorithms for analyzing price patterns and predicting stock prices and index changes. Most stock traders nowadays depend on Intelligent Trading Systems which help them in predicting prices based on various situations and conditions, thereby helping them in making instantaneous investment decisions.

Stock Prices are considered to be very dynamic and susceptible to quick changes because of the underlying nature of the financial domain and in part because of the mix of known parameters (Previous Days Closing Price, P/E Ratio etc.) and unknown factors (like Election Results, Rumors etc.)

An intelligent trader would predict the stock price and buy a stock before the price rises, or sell it before its value declines. Though it is very hard to replace the expertise that an experienced trader has gained, an accurate prediction algorithm can directly result into high profits for investment firms, indicating a direct relationship between the accuracy of the prediction algorithm and the profit made from using the algorithm.

1.2 Motivation behind the Project

In this paper, we discuss the Machine Learning techniques which have been applied for stock trading to predict the rise and fall of stock prices before the actual event of an increase or decrease in the stock price occurs. In particular the paper discusses the application of Support Vector Machines, Linear Regression, Prediction using Decision Stumps, Expert Weighting and Online Learning in detail along with the benefits and pitfalls of each method. The paper introduces the parameters and variables that can be used in order to recognize the patterns in stock prices which can be helpful in the future prediction of stocks and how Boosting can be combined with other learning algorithms to improve the accuracy of such prediction systems.

Note: The main goal of the project was to study and apply as many Machine Learning Algorithms as possible on a dataset involving a particular domain, namely the Stock Market, as opposed to coming up with a newer (and/or better) algorithm that is more efficient in predicting the price of a stock.

1.3 Overview of the Document

In Section 2 we try to briefly cover the background which is essential to the study of the domain of financial prediction systems. In Section 3 we discuss the results obtained from the application of the algorithms described in Section 1.2. Section 4 presents the concluding remarks for the experiment. Section 5 and 6 cover the Software Tools used and a list of research papers that were referenced (and that might serve as further reading material for those who are interested in this topic and want to explore it further.)

2. Background

2.1 Stock Prediction in Detail

In practice, there are 2 Stock Prediction Methodologies:

Fundamental Analysis: Performed by the Fundamental Analysts, this method is concerned more with the company rather than the actual stock. The analysts make their decisions based on the past performance of the company, the earnings forecast etc.

Technical Analysis: Performed by the Technical Analysts, this method deals with the determination of the stock price based on the past patterns of the stock (using time-series analysis.)

When applying Machine Learning to Stock Data, we are more interested in doing a Technical Analysis to see if our algorithm can accurately learn the underlying patterns in the stock time series. This said, Machine Learning can also play a major role in evaluating and forecasting the performance of the company and other similar parameters helpful in Fundamental Analysis. In fact, the most successful automated stock prediction and recommendation systems use some sort of a hybrid analysis model involving both Fundamental and Technical Analysis.

The Efficient Market Hypothesis (EMH)

The EMH hypothesizes that the future stock price is completely unpredictable given the past trading history of the stock. There are 3 types of EMH's: strong, semi-strong, and weak form. In the weak EMH, any information acquired from examining the stock's history is immediately reflected in the price of the stock.

The Random Walk Hypothesis

The Random Walk Hypothesis claims that stock prices do not depend on past stock prices, so patterns cannot be exploited since trends to not exist.

With the advent of more powerful computing infrastructure (hardware and software) trading companies now build very efficient algorithmic trading systems that can exploit the underlying pricing patterns when a huge amount of data-points are made available to them. Clearly with huge datasets available on hand, Machine Learning Techniques can seriously challenge the EMH.

Indicator Functions

We now take a brief look at the attributes and indicators that are normally used in the technical analysis of stock prices:

Indicators can be any of the following:

Moving Average (MA) : The average of the past n values till today.
Exponential Moving Average (EMA) : Gives more weightage to the most recent values while not discarding the older observation entirely.
Rate of Change (ROC) : The ratio of the current price to the price n quotes earlier.
n is generally 5 to 10 days.
Relative Strength Index (RSI): Measures the relative size of recent upward trends against the size of downward trends within the specified time interval (usually 9 – 14 days).

For this Project, the EMA was considered as the primary indicator because of its ability to handle an almost infinite amount of past data, a trait that is very valuable in time series prediction (It is worth noting that the application of other indicators might result in better prediction accuracies for the stocks under consideration).

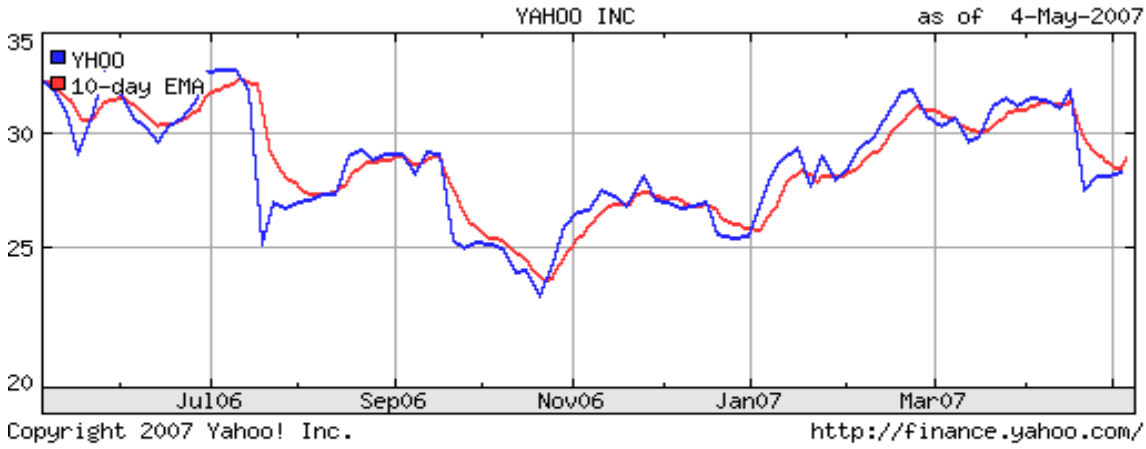
EMA (t) = EMA (t-1) + alpha * (Price (t) - EMA (t-1))

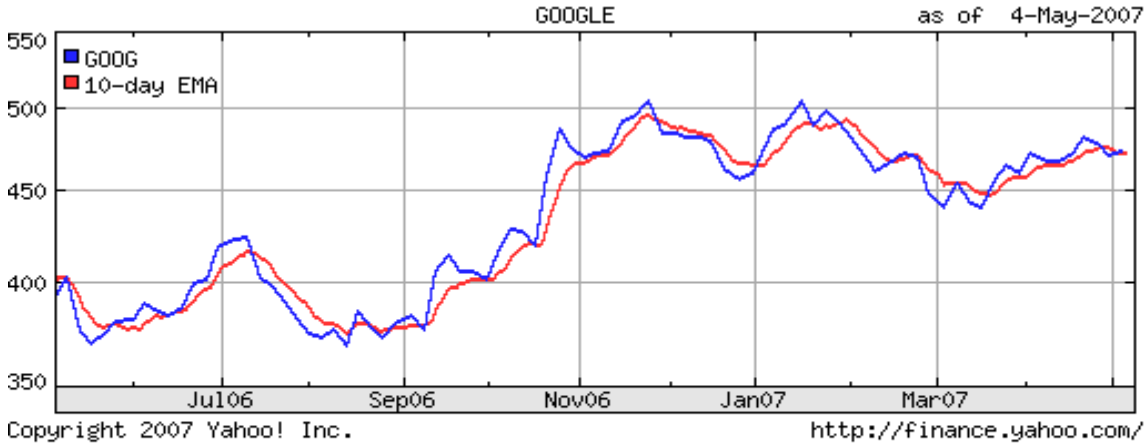
Where, alpha = 2/ (N+1), Thus, for N=9, alpha = 0.20

In theory, the Stock Prediction Problem can be considered as evaluating a function F at time T based on the previous values of F at times t-1,t-2,t-n while assigning corresponding weight function w at each point to F.

$F(t) = w1*F(t-1) + w2*F(t-2) + \ldots + w*F(t-n)$

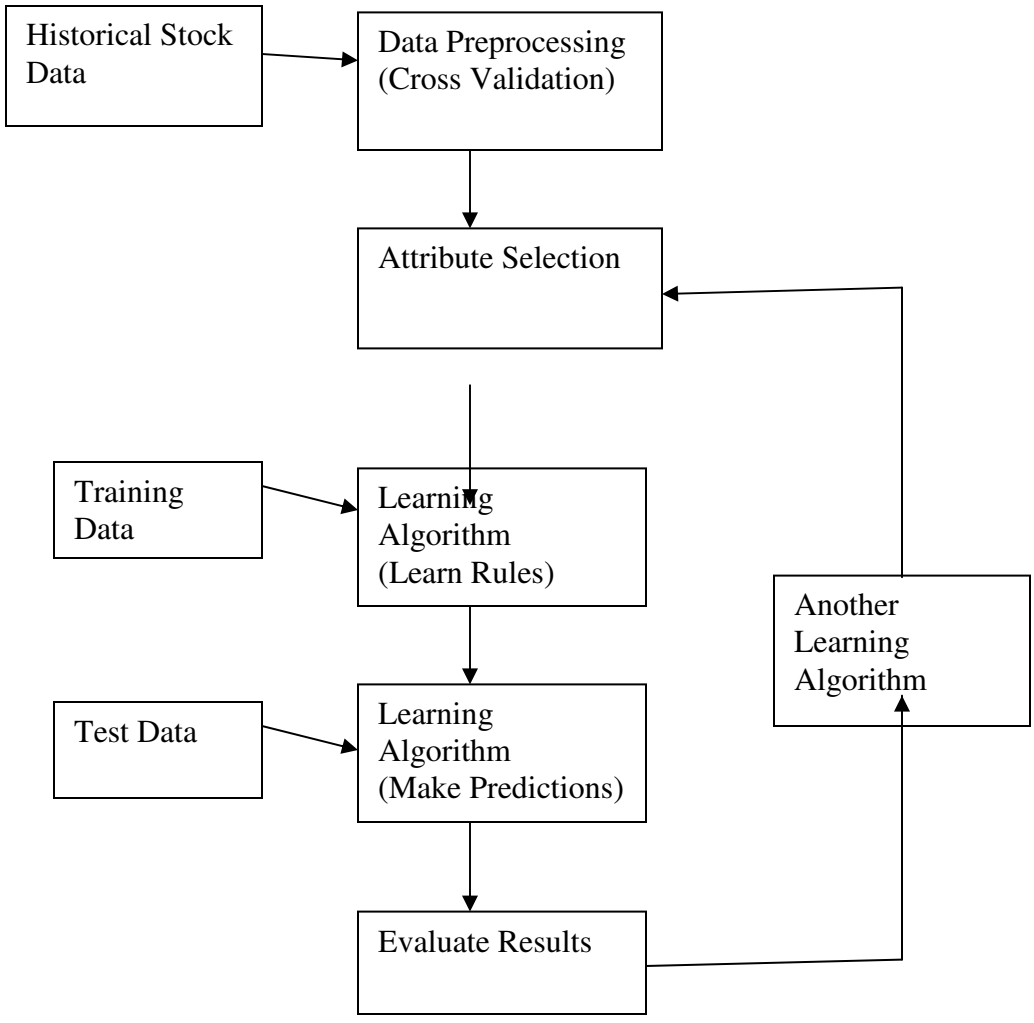
The technical analysis charts below show how the EMA models the actual Stock Price.





2.2 The Learning Environment

The Weka and YALE Data Mining Environments were used for carrying out the experiments. The general setup used is as follows:



Since the attribute space we are operating on consists of a very limited number of attributes (< 10) the Attribute Selection step can be skipped for some of the Machine Learning methods.

2.3 Preprocessing the Historical Stock Data

For this experiment, the historical data was downloaded from the yahoo finance section. In particular, the stock prices of two companies were studied, namely Google Inc. (GOOG) and Yahoo Inc. (YHOO)

The dataset available has the following attributes:
Date Open High Low Close Volume Adj. Close

Intuitively, based on the EMH, the price of the stock yesterday is going to have the most impact on the price of the stock today. Thus as we go along the time-line, data-points which are nearer to today's price point are going to have a greater impact on today's price. For a time-series analysis we can take the Date as the X-Axis with integer values attached to each date, such that the most recent Date Tag in the dataset gets the highest value and the oldest Date Tag gets the lowest value.

We add one more attribute to the above attributes, this attribute will act as our label for predicting the movements of the stock price. This attribute will be called "Indicator" and will be dependent on the other available attributes. For our experiments we use the EMA (Exponential Moving Average) as the indicator function.

3. The Machine Learning Techniques

In this section we evaluate the results generated on applying different learning algorithms.

3.1 Decision Stump

On applying a simple Decision Stump to predicting the EMA, we found the following results:

Correlation coefficient	0.8597
Mean absolute error	46.665
Root mean squared error	57.8192
Relative absolute error	46.8704 %
Root relative squared error	50.9763 %
Total Number of Instances	681

3.2 Linear Regression

On applying Simple Linear Regression (with only numeric attributes taken under consideration) the following results were obtained while predicting the EMA

Correlation coefficient	0.9591
Mean absolute error	12.9115
Root mean squared error	32.0499
Relative absolute error	12.9684 %
Root relative squared error	28.2568 %
Total Number of Instances	681

3.3 Support Vector Machines

Using C-Class Support Vector Machines which use RBF Kernels with the Cost Parameter C ranging from 512 to 65536, the accuracy in predicting the Stock Movement was as follows:

Root mean square error: 0.485 +/- 0.012
Accuracy: 60.20 +/- 0.49%

3.4 Boosting

The AdaBoostM1 Algorithm was applied to the DataSet after applying the C-SVC Algorithm. The results show a significant boost with respect to the Accuracy.

Root mean squared error: 0.467 +/- 0.008
Acuracy: 64.32% +/- 3.99%

The following confusion matrix was extracted from the output of the YALE Program (after applying a combination of C-SVC and AdaBoostM1)

True: 1 -1
1: 37 9
-1: 234 401
False Positive: 23.400 +/- 2.417

True: 1 -1
1: 37 9
-1: 234 401
False Negative: 0.900 +/- 1.221

True: 1 -1
1: 37 9
-1: 234 401
True Positive: 40.100 +/- 1.513

True: 1 -1
1: 37 9
-1: 234 401
True Negative: 3.700 +/- 2.410

3.5 Stock Prediction based on Textual Analysis of Financial News Articles

Nowadays, a huge amount of valuable information related to the financial market is available on the web. A majority of this information comes from Financial News Articles, Company Reports and Expert Recommendations (Blogs from valid sources can also act as a source of information) Most of this data is in a textual format as opposed to a numerical format which makes it hard to use. Thus the problem domain can now be viewed as one that involves Mining of Text Documents and Time Series Analysis concurrently.

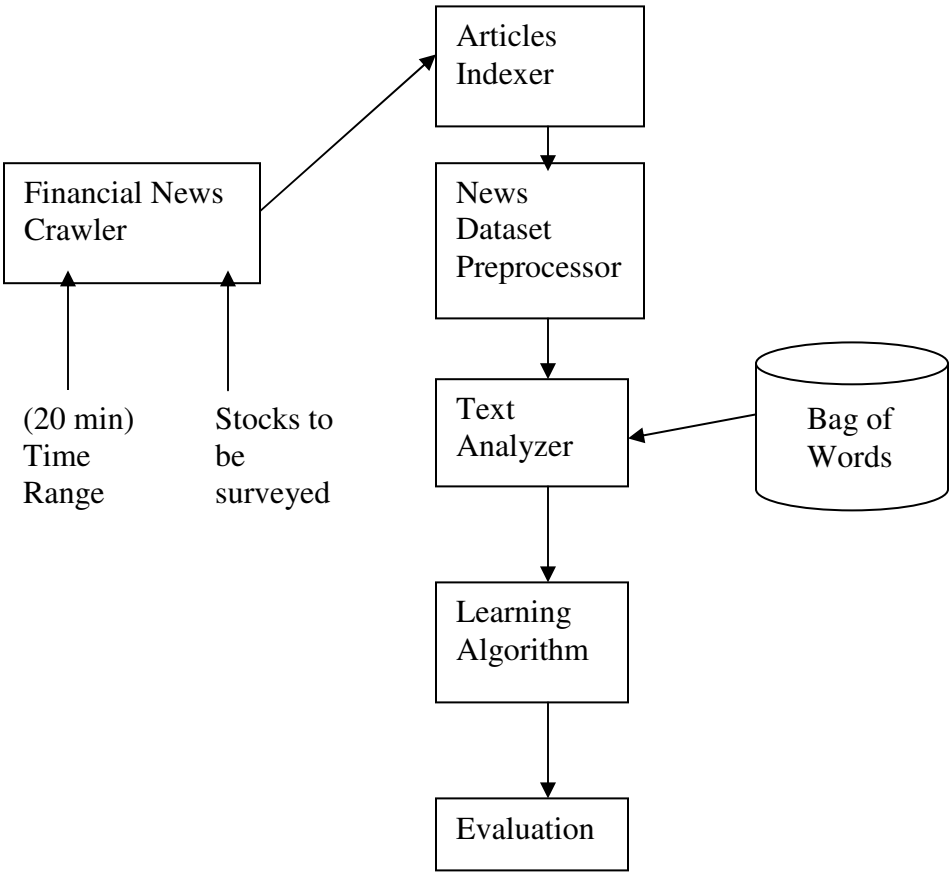
One method which has been used involves defining the news impact on a particular stock: Positive, Negative, and Neutral.

A news is considered to have a positive impact (or negative impact) if the stock price rises (or drops) significantly for a period, after the news story has been broadcasted. If the stock price does not change dramatically after the news is released, then the news story is regarded as neutral.

Another method which we study in this paper relates to detecting and determining patterns in the news articles which correspond directly to a rise or fall in the stock price. The general architecture is as follows:

A crawler continuously crawls news articles and indexes them for a particular stock portfolio. The learning environment requests the news since the last T minutes from the indexer. The learning environment consists of several base learners which look for specific information in the text document (i.e. patterns like “profits rise” inside a just released news article, or “share prices will go down” on the blog of a veteran Wall Street Trader/Speculator etc.). A Bag-Of-Words consisting of Positive Prediction Terms and Negative Prediction Terms and Phrases is used by the learning environment. Each time a word/phrase from the Positive Prediction Term occurs in a particular news article, a PostiveVote is assigned to the article.

The Diagram below shows the general architecture of such a system:



As can be seen, this method is very crude in making an accurate prediction.
To enhance the predictions, more weightage can be assigned to articles which come from credible sources. Also, more weightage can be assigned to a news Headline which contains a Positive Prediction Term or a Negative Prediction Term.
Boosting can then be applied to the base learners to see if the accuracy can be boosted further.

3.6 A brief discussion on Applying Expert Weighting to Stock Prediction

Star Analysts

Get **Star Analysts** for: GO

This is a list of top research analysts based on the accuracy of earnings estimates on GOOG, according to StarMine. Analysts that appear here are limited to those covering GOOG for a significant period of time. Learn More.

Total Ranked Analysts: 31

EPS ACCURACY FOR GOOG - Trailing Two Fiscal Years and Four Quarters

Top-Ranked Analysts	GOOG	Overall	Research Reports
Westerfield, Leland BMO Capital Markets	★★★★★		
Wolk, Marianne Susquehanna Financial Group	★★★★★	★★★★★	
Rohan, Jordan RBC Capital Markets	★★★★★		
Jain, Pratik First Global Stockbroking Ltd.	★★★★★		
Squali, Youssef Jefferies & Co.	★★★★★	★★★★★	
Garcia, Denise A. G. Edwards & Sons, Inc.	★★★★★		
Quarles, Christa Thomas Weisel Partners	★★★★☆	★★★★☆	
Brown, Derek Cantor Fitzgerald	★★★★☆	★★★★☆	

As shown in the table above, we are given the opinions of stock market experts (someone with an opinion, not necessarily the right one!) as input. At each round, we make our

prediction based on the predictions of the experts. For the subsequent rounds, we increase the weights of those experts who predicted the stock correctly and decrease the weights of those experts who were not correct in their predictions. (Another, often used variation to this method of expert weighting is to completely disregard those experts who were incorrect in their previous round prediction, this might yield in lower efficiencies (since even an expert is bound to make mistakes))

Thus the Expert Weighting Algorithm can be described as follows:

Given: A vector E = { e1, e2,….eN} of Stock Market Experts and their predictions.

Assign W(e(i)) = 1 For Each Expert e(i).
For Round t in 1…T
Make a Prediction based on the Weighted Majority Algorithm.
*For experts who made a correct prediction W(e(i))(t) = 2*W(e(i))(t-1)*
*For experts who made an incorrect prediction W(e(i))(t) = ½ * W(e(i))(t-1)*

Store the Expert Ratings for future weight assignments.

The topic of expert weighting based on expert opinions can be considered as a hybrid technique with influences from both the Fundamental Analysis and Technical Analysis domains since the experts make their opinions based on the principles of Fundamental Analysis and out expert weighting algorithm uses that data to do a technical analysis.

It should be noted that the methods described in 3.5 and 3.6 can be considered as a hybrid combination of Online Learning and Weighted Majority Algorithms because of the inherent characteristics like

1) Fetching Information One Piece at a time for a given time period.
2) Decisions are based on the past performance without knowing the future.
3) Adapt and learn as we go further.

4. Conclusion

Of all the Algorithms we applied, we saw that only Support Vector Machine combined with Boosting gave us satisfactory results. Linear Regression gave lower mean squared errors while predicting the EMA pattern.

Another technique which looks promising but which we did not cover the evaluation of was Expert Weighting. More recently, the linguistic analysis of Financial News Results to predict stocks has been a topic of extensive study.

The choice of the indicator function can dramatically improve/reduce the accuracy of the prediction system. Also a particular Machine Learning Algorithm might be better suited to a particular type of stock, say Technology Stocks, whereas the same algorithm might give lower accuracies while predicting some other types of Stocks, say Energy Stocks.

Moreover, we should also note that while applying the Machine Learning Algorithms for Technical Analysis, we assumed that the effect of the Unknown Factors (Election Results, Rumors, Political Effects etc.) was already embedded into the historical stock pattern. Commercial Trading systems might have a more sophisticated mechanism for taking the unknowns into account.

While we studied the algorithms discretely, more often than not, a hybrid algorithm is used for stock prediction. For instance an Algorithmic Trading System might involve a 3-tier architecture with SVM's and Boosting at the bottom, an Online Algorithm (For instance an Expert Weighting scheme that we discussed in section 3.6 as the middle layer and Textual Analysis of Stock Market News, Financial Reports as the top layer to make predictions.

As a result of the research conducted for this project, a subset of the algorithms discussed above have been implemented at the following web address:

http://www.stoocker.com

5. Tools//DataSets

Tools

LibSVM
http://www.csie.ntu.edu.tw/~cjlin/libsvm/

YALE (Yet another Learning Environment)
http://rapid-i.com/content/view/26/82/

WEKA
http://www.cs.waikato.ac.nz/ml/weka/

DataSets

Historical Stock Data available from http://www.finance.yahoo.com

Financial Stock Data also available for non-commercial purpose at:
http://www.econ.ubc.ca/whistler/325/dataset.htm

The most commonly available attributes are:
Date Open High Low Close Volume Adj. Close

6. References and Further Reading

Forecasting stock market movement direction with support vector machine
http://madis1.iss.ac.cn/madis.files/pub-papers/c&or-hw-hnw-04-1.pdf

On Developing a Financial Prediction System: Pitfalls and Possibilities
http://www.smartquant.com/references/NeuralNetworks/neural20.pdf

Prediction of Stock Market Index Changes
http://citeseer.ist.psu.edu/cache/papers/cs/129/ftp:zSzzSzftp.cs.bilkent.edu.trzSzpubzSztech-reportszSz1992zSzBU-CEIS-9201.pdf/sirin93prediction.pdf

J Moody, M Saffell, Learning to Trade via Direct Reinforcement, IEEE Transactions on Neural Networks, Vol. 12, No 4, July 2001.
http://www.cs.ucsd.edu/~dboswell/PastWork/Moody01LearningToTradeViaDirectReinforcement.pdf

AUTOMATED TRADING WITH BOOSTING AND EXPERT WEIGHTING
http://www.andromeda.rutgers.edu/~jmbarr/NYComp/CreamerEEA.pdf

Forecasting Stock Prices Using Neural Networks
http://www.andrew.cmu.edu/user/wyliec/project.pdf

Using Neural Networks to Forecast Stock Market Prices
http://people.ok.ubc.ca/rlawrenc/research/Papers/nn.pdf

MACHINE LEARNING IN COMPUTATIONAL FINANCE
http://www.cs.rpi.edu/~magdon/students/boyarshinov_victor/boyarshinov_PhDthesis.pdf

The Predicting Power of Textual Information on Financial Markets
http://www.comp.hkbu.edu.hk/~cib/2005/Jun/iib_vol5no1_article1.pdf

APPLICATION OF MACHINE LEARNING TO SHORT-TERM EQUITY RETURN PREDICTION
http://publish.uwo.ca/~jnuttall/cooper.pdf

Foreign Exchange Trading using a Learning Classifier System

http://www.cems.uwe.ac.uk/lcsg/reports/uwelcsg05-007r.pdf

Techniques and Software for Development and Evaluation of Trading Strategies
http://www.cs.umu.se/~thomash/reports/phdthesis.pdf

Data Mining for Prediction Financial Series Case
http://szemke.math.univ.gda.pl/zemke2003PhD.pdf

Reinforcement Learning for Optimized Trade Execution
http://www.cs.ualberta.ca/~sutton/kearnstradeexecution.pdf

Vatsal H. Shah | *vatsals@vatsals.com* **|** *vhs212@nyu.edu*
Foundations of Machine Learning | *Spring 2007*
Dr. Mehryar Mohri
Courant Institute of Mathematical Science
New York University