Stock Market Analysis

Department of Electrical Engineering and Computer Science
University of Central Florida

Sai Lalitha R
sailalitha@knights.ucf.edu
Graduate Student

Nishit Prasad
nshtpd@knights.ucf.edu
Graduate Student

Shruti Kamtekar
shruti.kamtekar@knights.ucf.edu
Graduate Student

Abstract

Stock market forecasting using machine learning algorithms is receiving attention among researchers & traders, but remains challenging due to random stock dynamics. In this project, we develop a novel prediction tool using comprehensive technical analysis of time series data to extract the up/down trend of the stock for next few weeks. We specifically extract 27 features from eight month time series data and utilize logistic regression & support vector machines training algorithms for classification with $\sim 65\%$ accuracy. In addition to trend prediction, we also conducted risk analysis by categorizing the stocks based on volatility, momentum & market indicators using the respective feature weights.

1 Introduction

Anticipating the stock value trend by interpreting the acceptable disordered market data using machine learning algorithm has been an interesting field for many researchers. According to the Efficient Market Hypothesis (EMH) an investment theory, states that it is difficult to "beat the market" since stock market efficiency causes existing share prices to constantly assimilate and reflect all relevant information. As indicated by the EMH, stocks trade at their fair value on stock exchanges, making it unthinkable for investors to either buy underestimated stocks or offer stocks for inflated costs. Short term and seasonal trend can contradict EMH, momentum in the stock price for few days adds predictability to stock over short term and seasonal variation such as holiday quarter can add to long term prediction.

When predicting stock price direction, investors commonly utilize of the three approaches. The first is the fundamental analysis, which looks at the monetary elements that drive the stock prices (e.g a company's financial statements such as the balance sheet or income statement). The second approach is to use Sentiment Analysis that uses sentiment from news headlines related to the company, public reviews etc. The third approach is technical analysis, that anticipates the stock price based on the past patterns of the price and volume of the stock using time series analysis.

Our project focuses on short term prediction of the stocks using technical indicators. We categorize the stocks in various sectors such health care, technology,capital goods etc. Our main agenda is to observe which features is more significant in a sector and how a stock depends on that feature to make the final prediction of the stock trend.

2 Related Work

Predicting stock market trend has become major topic of research among researches. There are many different way in which researchers have tried to predict the stock market trend. In the paper presented by Bryce Taylor [**Bryce, 2013**] i.e., Machine Learning Techniques for Stock Prediction, used various algorithms for stock prediction and found that support vector machine

along with boosting gave better results. Authors Duc Duong, Toan Nguyen and Minh Dang [**Doung, 2016**] use financial news articles to make stock prediction by reading all the HTML new articles for companies in VN30 Index and using SVM with prediction of 73% for dataset set of 1287 and 67.6 % for sample size of 1884 articles. Artificial neural networks (ANN) has been used for modeling financial time series [**Kim, 2000; Kohzadi, 1996**].

We focus on predicting the stock trend using SVM and Logistic regression for 8 categories of top 80 stocks based on market capital and try to predict the stock trend for next few weeks. Our project focuses on understanding which technical feature is more relevant for a category & stock using feature weights and determine the right mix of volatile, short-term and long-term safe stocks.

3 Our Approach

We start with the time-series data of given stock obtained from Yahoo Finance. The data consists of daily open, close, high, low & volume day. Figure 1 shows data for AAPL from May-Nov 16 with daily price & volume vertical axes. The colour of stick is either black/red indicating up/down trends respectively.

3.1 Datasets

The stock financial data is pulled from Yahoo! Finance which is a part of Yahoo!'s network that provides financial stocks data, news and reports. This media property provide various tools for personal finance management and analysis. We are using nine months of data from November 1, 2015 to July, 15 2016. The ranges of stock data that we considered are based on the market capital. We considered top 80 stocks sorted in the order of highest to lowest market capital value. These stocks or companies belong to one or the other industry sectors. The following are the industry sectors that are prevalent on a major scale:

The 80 stocks that we have considered based on the highest market capital, are non-uniformly segregated with their respective industry sectors. In Table 1, we see 8 sectors, namely, Capital Goods, Consumer Goods, Finance, Health Care, Miscellaneous, Public Utilities, Technology and Transportation. Certain group of stocks out of the 80, belong to any one of the aforementioned sectors. For example, Apple Inc. belongs to Technology sector, Walgreens Alliance Inc. to Health Care, etc. Table 2 shows some of the stocks, i.e., companies that we have considered with their respective industry sectors.

Capital Goods	Consumer Goods
Finance	Health Care
Miscellaneous	Public Utilities
Technology	Transportation

Table 1: Industry sectors

Stock List	
Stock Name	**Industry Sector**
Tesla Motors, Inc.	Capital Goods
PACCAR Inc.	Capital Goods
Illumina, Inc.	Capital Goods
The Kraft Heinz Company	Consumer Goods
Mondelez International, Inc.	Consumer Goods
Monster Bvg. Corporation	Consumer Goods
Amazon.com, Inc.	Consumer Goods
...	...
CME Group Inc.	Finance
Fifth Third Bancorp	Finance
...	...
Amgen Inc.	Health Care
Gilead Sciences, Inc.	Health Care
...	...
Apple Inc.	Technology
Google	Technology
Microsoft Corporation	Technology
..	..
..	..

Table 2: Stocks with respective industry sectors

Each stock has a four-letter stock-code word that is used for implementation in order to get financial data for a particular date range. Since we used Python as the core language for our implementation, there is library called pandas that has predefined functions to pull financial data. These functions take stock-code, followed by stock source as 'yahoo' and then the date range.

3.2 Building Feature matrix

The first task is to eliminate the noise in the data by fitting moving averages. Two classes of moving averages are popular in trading industry,

Figure 1: Candlestick plot of raw data

simple moving average (SMA) and exponential moving average (EMA). While SMA is a simple average over a fixed number of days, EMA provides weighted average by giving more significance to the prices in most recent days.

In this paragraph, we will discuss how moving averages serve as technical indicators. Figure 1 shows SMA5, EMA5 & SMA50 curves which represents SMA/EMA for 5 days (short-term) & SMA for 50 days (long-term) respectively. The points where SMA5 and SMA50 crossover are the key points which denote buy/sell signals. Soon after crossover, if the SMA5 curve falls behind SMA50 then it is considered as a sell signal. Likewise if the SMA5 curve goes above SMA50, it is taken as a buy signal. The equivalent EMA crossover points are captured by MACD feature which is also an important indicator used by the traders.

For this project, we have extracted 27 features to comprehensively capture the stock dynamics [**Xinjie, 2014**]. The acronyms and description of each feature is described in Appendix A. We divided these features into the following 4 different categories:

(i) Momentum based: These measure the rate of rise or fall in stock prices. **Features:** MOM, WILLR, RSI

(ii) Trend based: These indicate the trend of the stock by taking average price over a period. Crossovers of EMA noted by MACD is used to generate buy/sell signals. **Features:** TSF, SMA, EMA, MACD, RSI, OBV

(iii) Volatility based: These indicators monitor changes in market price and compare them to historical values. In case the price crosses the upper bound of the range, in which it should be present per the trends in historical data, a sell signal is generated. Conversely, if the price falls below the lower bound of the range, a buy signal is generated. **Features:** ROCR, BUPPER, BLOWER, BMIDDLE

(iv) Market based: These are a series of technical indicators used by traders to predict the direction of the major financial indexes. They follow the market like S&P and NASDAQ and fluctuate with them. **Features:** Features of SandP that are added to stock like adxSnP, blowerSnP etc.

In order to build the feature matrix, all the 27 features are gathered to form a matrix with features as the columns and date as the rows. For each stock, we also add S&P 500 features which captures overall trend of market. In total, combining 27 features of the stock and 27 features of the S&P 500 results in 54 features. Therefore the feature matrix is a MxN matrix where N is 54 i.e., total number of features and M is number of trading days. Feature matrix is further normalized so that there are no non zero elements (NaN) and all the number are in the same range of 0 to 1.

3.3 Building Y matrix

The stock data is labeled with up or down indicated as +1 and -1 respectively. There are many approaches to derive the training labels from the dataset. For e.g., training labels can be created by predicting the next day trend or next 3-day average price trend. For predicting next day trend,

$$Y(t) = \begin{cases} 1 & \text{if Price(t+1)} > \text{Price(t)} \\ -1 & \text{if Price(t+1)} < \text{Price(t)} \end{cases}$$

For predicting next 3-day average price trend,

$$Y(t) = \begin{cases} 1 & \text{if SMA(t+3)} > \text{SMA(t)} \\ -1 & \text{if SMA(t+3)} < \text{SMA(t)} \end{cases}$$

where $\mathbf{Y(t)}$ is the training label for the data and SMA(t) is the Simple Moving Average for the data at time t. It is important to note that feature matrix $X_i(t)$ do not see future data, but

Y(t) is calculated based on future data (1 or 3 days into future). In this project, next 3-day average price trend is considered to predict for all the stocks.

In addition to binary class labelling, we further extended the model to multi-class labelling using five different classes for stock sentiment like: Very Bullish (1), Bullish(2), Neutral (3), Bearish (4) and, Very Bearish (5).

3.4 Feature selection

As introduced in Sec. 3.1, there are 54 features in the initial feature bag. To know if all the features are important and if only subset of features can be used without losing prediction accuracy we perform feature selection. In the present project, a feature selection algorithm, i.e., random forest feature selection is applied. In Random forest feature selection: Random forest model can provide scores for features by ranking. The larger the score is, the more important is the feature. Top 30% of the features are selected as per ranking and are used to train the model. All the features are categorized into four different types based on what they indicate.

Comparison among the results of all the stocks shows that, the feature combination is different for different stocks. For example, Figure 2 show the selected features for 10 different technology stocks. It indicates that AAPL(Apple) stock is mostly based on momentum based indicators to decide price change. This means that AAPL stock can be bought for short-term gains and as soon as the momentum starts to fade, it is advisable to sell this stock. On the other hand, CSCO(Cisco) stock is dependent on Market indicator i.e., S&P. It means CSCO follows the market trend and is better holding for long term. For the case of Applied Materials (AMAT), volatility indicators are dominant factors in determining the price trend. It means there is both risk and reward associated with owning this stock.

3.5 Model training using Logistic regression and SVM

For training the model we are using feature matrix X and Y matrix mentioned in section 3.2

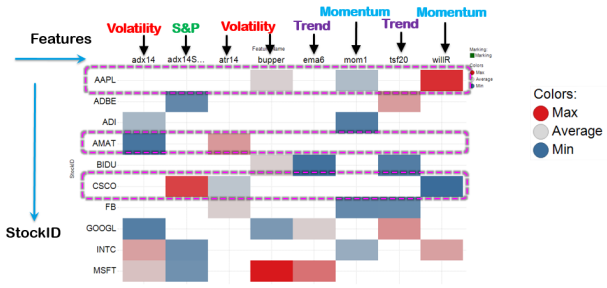

Figure 2: Candlestick plot of raw data

and 3.3 of the paper respectively. X and Y matrix are divided into a 80-20 ratio, where 80% of the data is used for training and 20% for testing. To additionally break down the features that are more significant for category and stock, feature selection as mentioned in section 3.4 is applied on the feature matrix before splitting the feature matrix into 80-20 ratio.

To test the model we are using two machine learning algorithms logistic regression and support vector machine using RBF kernel. We compared two approaches in order to build a better model for predicting stock future effectively and efficiently. Logistic regression is one of the basic classification algorithm which is used for multi-class classification of the data. Regularized logistic regression is used to train the model as it can handle both dense and sparse inputs.

The general idea of non-linear SVM states that the original input space can always be mapped to higher dimensional feature space when the training set is separable. The classifying function for non-linear SVM is:

$$f(x) = \sum \alpha_i y_i \theta(x_i)^T \theta(x) + b$$

As SVM relies on the inner product, instead of calculating it kernel function is used. The kernel function for RBF(Radial Basis Function) network is:

$$K(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$$

where $gamma = \frac{1}{2\sigma^2}$ and C is optimization criterion to avoid overfitting.

In SVM algorithm, C and gamma are the two important parameters. In order to get the best combination of C and gamma we used 'Grid-

SearchCV' technique which picks the best parameter combination and avoids overfitting of the model. Training model with SVM is proved to be one of the most efficient techniques for stock market prediction.

After comparing LR and SVM algorithms, it is observed that SVM can train the model efficiently when passed with 54 features. It gives an average accuracy of $\sim$ 65% for 88 different category stocks.

3.6 Plotting Graphs

For graphical representation of the results and better understanding we are using mathplotlib [**?**] library of python. We are plotting graphs for each algorithm for all the stocks in one category. By plotting of graph of selected features can get better understanding of the features that are prominent in a single category of all stocks.

4 Experiments

Feature calculation and usage plays an important role in predicting the stock market. After building the feature and Y matrices Logistic Regression and SVM algorithms are used to train the model and test it. Multiple experiments are conducted to know the best features and avoid overfitting.

As mentioned, we have considered four scenarios in terms of number of features. The purpose of the experiments is to determine how many features do we really need to consider to get decent accuracy and avoid overfitting. The first case considers passing all 54 features into X-matrix (27 features of a given stock & 27 features of SnP), the second case considers passing only 27 features of given stock in X, the third case is selecting top 30% or 8 features (determined by randomized tree algorithm) of the given stock into X and finally the fourth case considers selecting top 30% features from both stock & SnP i.e., 16 features into X-matrix. Training and testing accuracies for 88 stocks are stored in an excel file for all the experiments. The names of files are mentioned below:

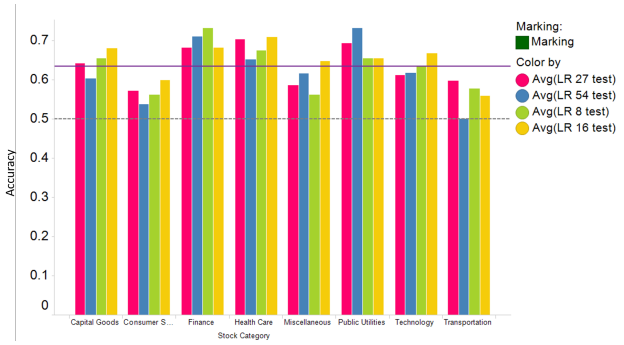

Figure 3: Accuracies of 88 stock categorized into 8 sectors using Logistic Regression algorithm

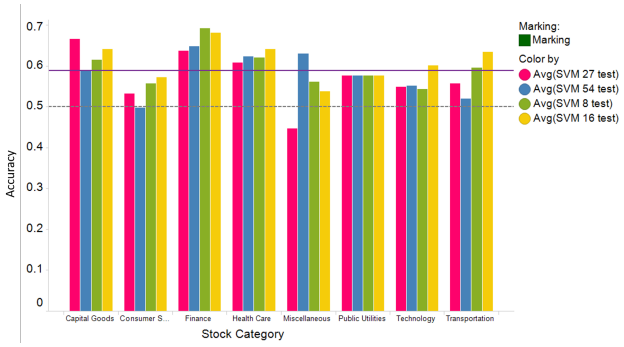

Figure 4: Accuracies of 88 stock categorized into 8 sectors using Support Vector Machine algorithm

$$LR \left\{ \begin{array}{c} \text{27 feat(Stock)} \\ \text{54 feat(Stock + SnP)} \\ \text{8 feat(Selected from 27)} \\ \text{16 feat(Selected from 54)} \end{array} \right\} LR.xslx$$

$$SVM \left\{ \begin{array}{c} \text{27 feat(Stock)} \\ \text{54 feat(Stock + SnP)} \\ \text{8 feat(Selected from 27)} \\ \text{16 feat(Selected from 54)} \end{array} \right\} SVM.xslx$$

Top selected features of 88 stocks for SVM after applying tree algorithm is stored in 'Selectedfeatures-SVM.csv' file.
All the plotted graphs for 88 stocks categorized into 8 sectors are stored in 'Graphs' folder.

Figure 3 shows accuracy of test set on all 88 stocks (complete list can be found in the results folder) using binary logistic regression analysis. In order to be concise, we plot the average accuracies of bunch of stocks based on the market sector they belong to. For example, 30 tech

stocks are grouped together under 'Technology' axis and the average accuracy is plotted. We find that LR predicts price trend in all sectors better than 60%. Figure 4 shows accuracy of test set on all 88 stocks using binary SVM with RBF kernel. We find that SVM also predicts price trend in all sectors better than 60%. The average line in each figure indicates "overall accuracy" of each model and are 64% and 61% respectively. We find that better accuracy is obtained when we use 16 features in X matrix, i.e., top 30% features selected from 27 features of stock and 27 features of SnP.

5 Conclusion

In this project, we use supervised learning techniques- Logistic Regression and SVM with RBF kernel in predicting the stock price trend of multiple (88) stocks over a time period of 6 months. Our findings show that our approach is feasible in providing good accuracy in prediction of stock prices in near future. Based on the time series data, we have extracted comprehensive 27 feature vectors and built train matrix using both stock data and financial index SnP. We find that an accuracy better than 60% is obtained using both LR and SVM methods with higher accuracy when top 30% features are considered in X matrix. In addition to uptrend/downtrend prediction, our tool also gives valuable insights into the technical indicators affecting the stock dynamics. Specifically, our tool can categorize the stock into either 'volatile' nature, 'market-driven', 'short-term momentum driven' or 'trend-based'. This analysis thus provides an indispensable tool for investors to better balance their portfolio by having right mix of stocks.

The model can be further bolstered by adding news sentiment, seasonal indicators and global economic indicators into the training model. This will require (i) gathering processing news articles that mention a given stock, (ii) adjust the sentiment according to season (ex: retailers are more bullish during fall holidays) and (iii) tune the model to include global issues (ex: foreign elections influencing trade of some commodities). When such indicators are included into training the model, we anticipate more accurate prediction of short-term stock movement. We plan to append this information into an already exhaustive list of technical indicators developed in this paper as a future extension to our model.

6 Appendix

Definitions and formulas of symbols are explained in detail in the Table 6 below. In the project, a library named '' (used extensively by financial traders) is used to get the values of the features.

In Table 6 Price(t) = Adj.Close(t)

References

Bryce Taylor, Applying Machine Learning to Stock Market Trading, 2013

Duong, Duc, Toan Nguyen, and Minh Dang, "Stock Market Prediction using Financial News Articles on Ho Chi Minh Stock Exchange." Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication. ACM, 2016.

Cheng, S. "A neural network approach for forecasting and analyzing the price-volume relationship in the Taiwan stock market." Master's thesis, National Jow-Tung University, Taiwan, ROC (1994).

Kim, Kyoung-jae, and Ingoo Han. "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index." Expert systems with Applications 19.2 (2000): 125-132.

Kohzadi, Nowrouz, et al. "A comparison of artificial neural network and time series models for forecasting commodity prices." Neurocomputing 10.2 (1996): 169-181.

Xinjie Di, Stock Trend Prediction with Technical Indicators using SVM, 2014

Table 3: Technical indicators used as features

Indicators	**Name**	**Description**	**Formula**
OBV	On Balance Volume	Relates trading volume to price change	OBV(t) = OBV(t-1) +/- Volume(t)
RSI	Relative Strength Index	Suggests the overbought and oversold market signal	Avg(PriceUp)/(Avg(PriceUP)+ Avg(PriceDown))*100 Where: PriceUp(t)=1*(Price(t)-Price(t-1))Price(t)- Price(t-1)>0; PriceDown(t)=1*(Price(t-1)-Price(t))Price(t)- Price(t-1)<0;
SMA	Simple Moving Average	Smoothen the curve	Sum(Adj.Close(n))/n Eg: SMA3 is Adj.close average for 3 days
EMA	Exponential Moving Average	Smoothen the curve by giving more weight to recent prices	EMA [today] = (Price [today] x K) + (EMA [yesterday] x (1 – K)) Where: K = 2 /(N + 1); N = the length of the EMA; Price [today] = the current closing price; EMA [yesterday] = the previous EMA value; EMA [today] = the current EMA value;
ATR	Average True Range	Shows volatility of the market	ATR(t)=((n-1)*ATR(t-1)+Tr(t))/n where Tr(t)=Max(Abs(High-Low), Abs(Hight-Close(t-1)), Abs(Low-Close(t-1));
MFI	Money Flow Index	Relates price with volume	100-(100/(1+Money Ratio)) where Money Ratio=(+Moneyflow/-Moneyflow); Moneyflow=Tp*Volume
ADX	Average Directional Index	Discover if trend is developing	Sum((+DI-(-DI))/(+DI+(-DI))/n; where DI is Directional Index
MOM	Momentum	Measures the change in price	Price(t)-Price(t-n)
CCI	Commodity Channel Index	Identifies cyclical turns in stock prices	CCI = (Typical Price - 20-period SMA of TP) / (.015 x Mean Deviation) Typical Price (TP) = (High + Low + Close)/3 Constant = .015
ROCR	Rate Of Change	Compute rate of change relative to previous trading intervals	2(Price(t)/Price(t-n))*100
outMACD	Moving Average Convergance Divergance	Use EMA to signal buy/sell	(EMA for 12 days – EMA for 26 days)
outMACDSignal	MACD signal	signal indicator	Calculate EMA of MACD series for 9 days

outMACDHist	Histogram	Gets histogram	(MACD Series – Signal Line) ; Signals generated based on crossovers between MACD and signal line
WILLR	Williams%R	Determines where today's closing price fell within the range on past 10- days	(highest-closed)/(highest-lowest)*100
TSF	Time Series Forecasting	Calculates the linear regression of 20 day price	Linear Regression Estimate with 20-day price.
TRIX	Triple Exponential Moving Average	Smooth the insignificant movements	TR(t)/TR(t-1) where TR(t)=EMA(EMA(EMA(Price(t)))) over n days period
bupper	Bollinger Upper Band	Volatility Indicator	Middle Band + (y * n-period standard deviation)
bmiddle	Bollinger Middle Band	Volatility Indicator	n-period moving average
blower	Bollinger Lower Band	Volatility Indicator	Middle Band - (y * n-period standard deviation)