International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 3

Stock Prediction System Based on Key Statistics for S&P 500 With Linear SVC

G. Saminath Krisna, Dr. R. Indra Gandhi
Department of Computer Science, GKM College
of Engineering and Technology, Chennai, Tamil Nadu, India

ABSTRACT

Previous research shows strong evidence that traditional regression- based predictive models face significant challenges in predictability tests due to uncertain models and unstable parameters. Recent studies introduce new, stable strategies to overcome these problems. Support Vector Clustering is a relatively new learning algorithm that has the desirable characteristics of the control of the decision function, the use of the kernel method, and the sparsity of the solution. In this paper, we present a theoretical and empirical framework to apply the Support Vector Machines strategy to predict the stock market. There are many factors like macro and microeconomic events that may influence the stock trend. For predicting the stock performance, Support Vector Machine is used to analyze the relationship between these factors. Our results suggest that support vector clustering is a powerful predictive tool for stock predictions in the financial market.

Keywords: *Stock prediction, predictive models, predictive algorithms and training data*

INTRODUCTION

Quantitative trading using artificial intelligence and machine learning algorithms is raising much interest recently. Machine learning algorithms leveraging Big Data and novel hardware can form the basis for effective stock price movement reasoning and artificial intelligent trading decision making. Machine learning algorithms are becoming better at various tasks from manual to cognition. This gives traders a wide range of new insights and opportunities to combine prior knowledge of financial market with non-observable information into trading strategies. These algorithms can accurately predict the fluctuation of stock price bring huge underlying income for fund managers and financial investors. Machine learning algorithms leveraging big data are able to form the basis for effective stock price prediction and trading decision making, so quantitative trading algorithms using machine learning and data mining technologies have been raising much interest these years. Consequently, a wide range of new insights and opportunities are provided for the traders to combine prior knowledge of finance with non-observable information.

REVIEW OF LITERATURE

The review of various works brings out various interesting facts. There are several studies that apply data mining and analysis tools to investigate the predictability of various financial series.

The reasons quoted for applying data mining tools are its ability to handle voluminous data and nontrivial extraction of implicit, previously unknown, and potentially useful information from financial data. The literature review point out the fact that results of prediction varies among different financial markets. The other interesting fact is that data mining tools are found to consistently outperform other statistical approaches. The literature suggests that data mining tools are likely to predict stock market price movements better when compared to other methods. . The most obvious advantage of the data mining techniques is that they can outperform the classical

statistical methods with 520% higher accuracy rate. Majority of the studies incorporated technical factors, intraday movements and macro-economic indicators as input factors for predicting the stock price movements. However, not many studies are concerned with including the global cues and its influence on predicting the stock index movements. Higher evolutionary computing tools like Random Forest and SVM, increased the efficiency of prediction from KNN and ANN with use of. About 60% of the articles used feed forward neural networks and recurrent networks. The literature review highlights that both statistical and nonstatistical measures were used to evaluate the efficiency of the data mining approaches adopted by researchers. The review of previous works showed that the earlier works undertaken are highly empirical and it is inferred that new research works in new markets at different time periods are likely to show new results and insights. The review also points out that not many studies have been undertaken in Indian and other emerging market economies. Also a comprehensive comparative analysis of the above markets is not undertaken using a data mining approach. The selection of tools initially started from ANN, and moved to Support Vector machines. The SVC (Support Vector Clustering) method was not used for prediction of stock market trend in many studies. Hence an attempt to use successful SVC and other data mining tools to evaluate the predictability of financial markets.

MAJOR STAGES

1) Data Acquisition
2) Feature extraction
3) Getting current data
4) Prediction

Stock prediction is one of the important applications of machine learning that has elevated in recent years. Stock prediction process depends upon number of factors like economic factors, rumours, investors sentiments, management quality fundamental statistics, etc and these factors influence the movement of stock price. There are four different phases in a stock prediction system, namely: Data acquisition, feature extraction, getting current data and prediction.

DATA ACQUISITION

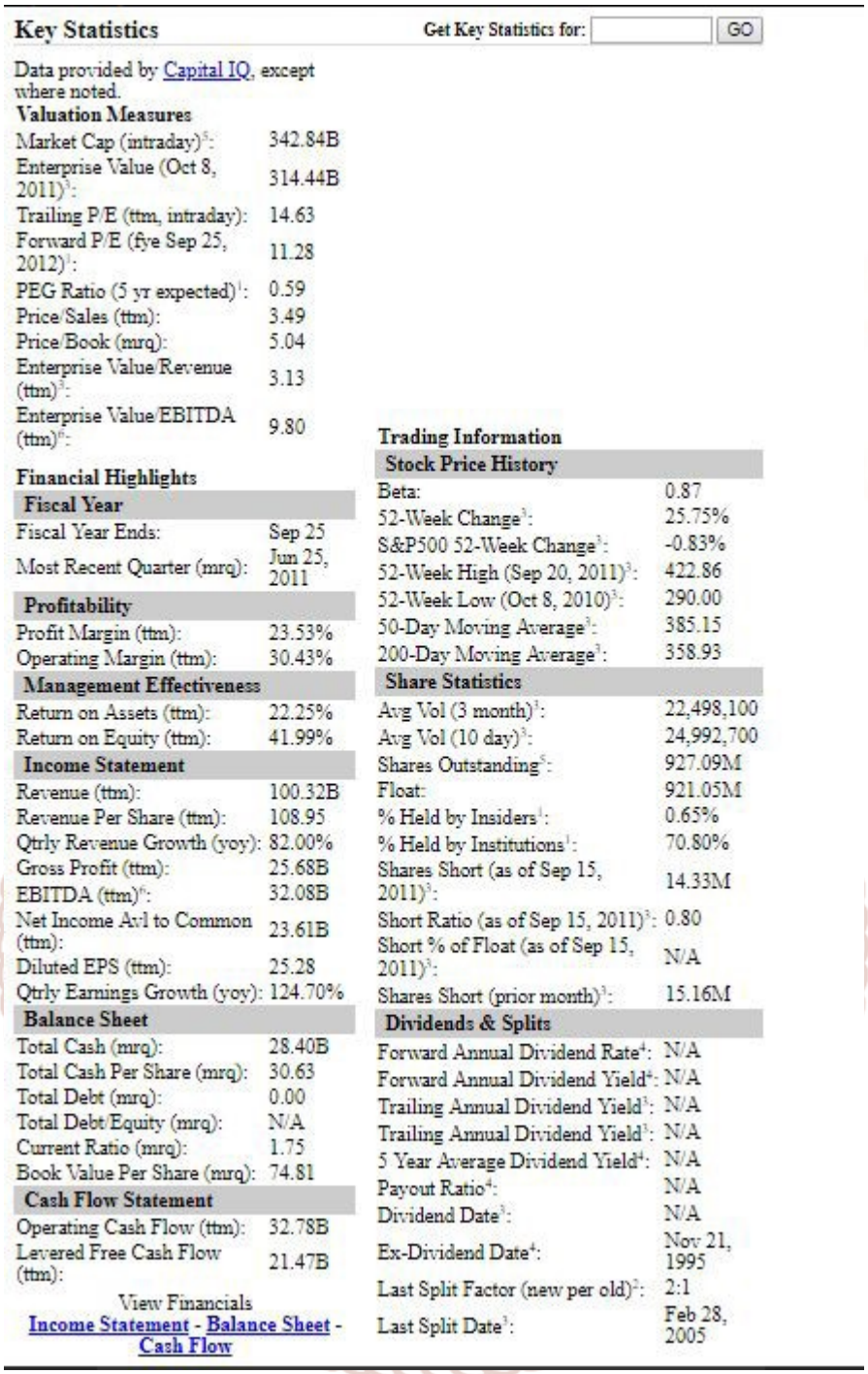
Data acquisition is the first stage is the stock prediction system. The data can be acquired either by parsing a stock quote website or by using a data providing service like quandl. The data acquired by parsing websites are stored as html files in the system and os walk is used to acquires the data within the web pages. The webpages are then analysed to find the required data. The acquired data is then stored in a csv (comma separated value) file. If the data is acquired from data providers like Quandl or Bloomberg then, the data comes by default in the csv format. So if you are using data providers to get your data, then data acquisition becomes much easier.

```
In [3]: df = pd.read_csv("key_stats.csv")
        df.head()
```

	Unnamed: 0	Date	Unix	Ticker	Price	stock_p_change	SP500	sp500_p_change	Difference	DE Ratio	...	Book Value Per Share	Cash Flow	Beta	Held by Insiders
0	0	2011-08-02 06:45:37	1.312293e+09	a	40.79	11.296044	1254.05	10.867009	0.429035	54.360	...	11.360	1.385000e+10	1.550	0.25
1	1	2013-05-14 00:28:58	1.368517e+09	a	43.04	17.435198	1650.34	45.901886	-28.466688	44.120	...	15.410	1.320000e+09	1.660	0.23
2	2	2013-09-06 06:09:52	1.378473e+09	a	47.68	30.095498	1655.17	46.328892	-16.233394	56.390	...	14.460	1.260000e+09	1.810	0.23
3	3	2004-04-13 04:20:10	1.081855e+09	aa	34.26	0.000000	1129.44	0.000000	0.000000	0.595	...	14.078	2.530000e+09	1.808	1.00
4	4	2004-09-09 16:15:45	1.094772e+09	aa	32.89	-3.998832	1118.38	-0.979246	-3.019586	0.561	...	14.049	1.780000e+09	1.676	1.00

Parsing data from websites

There are many frameworks and libraries available for parsing websites. Read the documentation of the website carefully before parsing it. Some websites do not allow developers to parse the websites without permission. The most common library for python to parse websites is beautiful-soup-4. There are also many other good libraries available for python which makes it easy to parse websites. The parsed websites can be stored in the system based on the system based on the ticker symbol and date.



Key Statistics Get Key Statistics for: GO

Data provided by Capital IQ, except where noted.

Valuation Measures

Market Cap (intraday)[5]:	342.84B
Enterprise Value (Oct 8, 2011)[3]:	314.44B
Trailing P/E (ttm, intraday):	14.63
Forward P/E (fye Sep 25, 2012)[1]:	11.28
PEG Ratio (5 yr expected)[1]:	0.59
Price/Sales (ttm):	3.49
Price/Book (mrq):	5.04
Enterprise Value/Revenue (ttm)[3]:	3.13
Enterprise Value/EBITDA (ttm)[6]:	9.80

Financial Highlights

Fiscal Year	
Fiscal Year Ends:	Sep 25
Most Recent Quarter (mrq):	Jun 25, 2011
Profitability	
Profit Margin (ttm):	23.53%
Operating Margin (ttm):	30.43%
Management Effectiveness	
Return on Assets (ttm):	22.25%
Return on Equity (ttm):	41.99%
Income Statement	
Revenue (ttm):	100.32B
Revenue Per Share (ttm):	108.95
Qtrly Revenue Growth (yoy):	82.00%
Gross Profit (ttm):	25.68B
EBITDA (ttm)[6]:	32.08B
Net Income Avl to Common (ttm):	23.61B
Diluted EPS (ttm):	25.28
Qtrly Earnings Growth (yoy):	124.70%
Balance Sheet	
Total Cash (mrq):	28.40B
Total Cash Per Share (mrq):	30.63
Total Debt (mrq):	0.00
Total Debt/Equity (mrq):	N/A
Current Ratio (mrq):	1.75
Book Value Per Share (mrq):	74.81
Cash Flow Statement	
Operating Cash Flow (ttm):	32.78B
Levered Free Cash Flow (ttm):	21.47B

View Financials
Income Statement - Balance Sheet - Cash Flow

Trading Information

Stock Price History	
Beta:	0.87
52-Week Change[3]:	25.75%
S&P500 52-Week Change[3]:	-0.83%
52-Week High (Sep 20, 2011)[3]:	422.86
52-Week Low (Oct 8, 2010)[3]:	290.00
50-Day Moving Average[3]:	385.15
200-Day Moving Average[3]:	358.93
Share Statistics	
Avg Vol (3 month)[3]:	22,498,100
Avg Vol (10 day)[3]:	24,992,700
Shares Outstanding[5]:	927.09M
Float:	921.05M
% Held by Insiders[1]:	0.65%
% Held by Institutions[1]:	70.80%
Shares Short (as of Sep 15, 2011)[3]:	14.33M
Short Ratio (as of Sep 15, 2011)[3]:	0.80
Short % of Float (as of Sep 15, 2011)[3]:	N/A
Shares Short (prior month)[3]:	15.16M
Dividends & Splits	
Forward Annual Dividend Rate[4]:	N/A
Forward Annual Dividend Yield[4]:	N/A
Trailing Annual Dividend Yield[3]:	N/A
Trailing Annual Dividend Yield[3]:	N/A
5 Year Average Dividend Yield[4]:	N/A
Payout Ratio[4]:	N/A
Dividend Date[3]:	N/A
Ex-Dividend Date[4]:	Nov 21, 1995
Last Split Factor (new per old)[2]:	2:1
Last Split Date[3]:	Feb 28, 2005

Data providers are corporations which sell data in exchange for money. There are lot of companies that sell data with different subscription models. Some of the well known websites for getting data are Quandl and bloomberg. The OHLC data can be acquired for free from many websites. But for this project we need the fundamental statistics of all the companies in s&p 500. For that we have to subscribe to any of the above mentioned data providers. They also have API (Application Programming Interface) and JSON (Javascript Object Notation) modes to get data other than CSV (Comma Separated Values).

FEATURE EXTRACTION

Munging is the process of clearing out the redundant information from the obtained data. If the data is obtained by parsing websites, then this process becomes much more complicated. The data obtained from data providers are usually pre processed and the redundant data is removed at the source. The important features are extracted in this stage. If we are using an API for getting data from the data provider, then we can specifically get the required data. The parsed data should also be stored in the csv or table format to process it using pandas. The pandas library

converts the csv file into a dataframe and the dataframe is used for further processing. SVC is a supervised form of learning. So the data has to be labeled as either outperforming or underperforming.

```
In [4]: df.head(5)
```

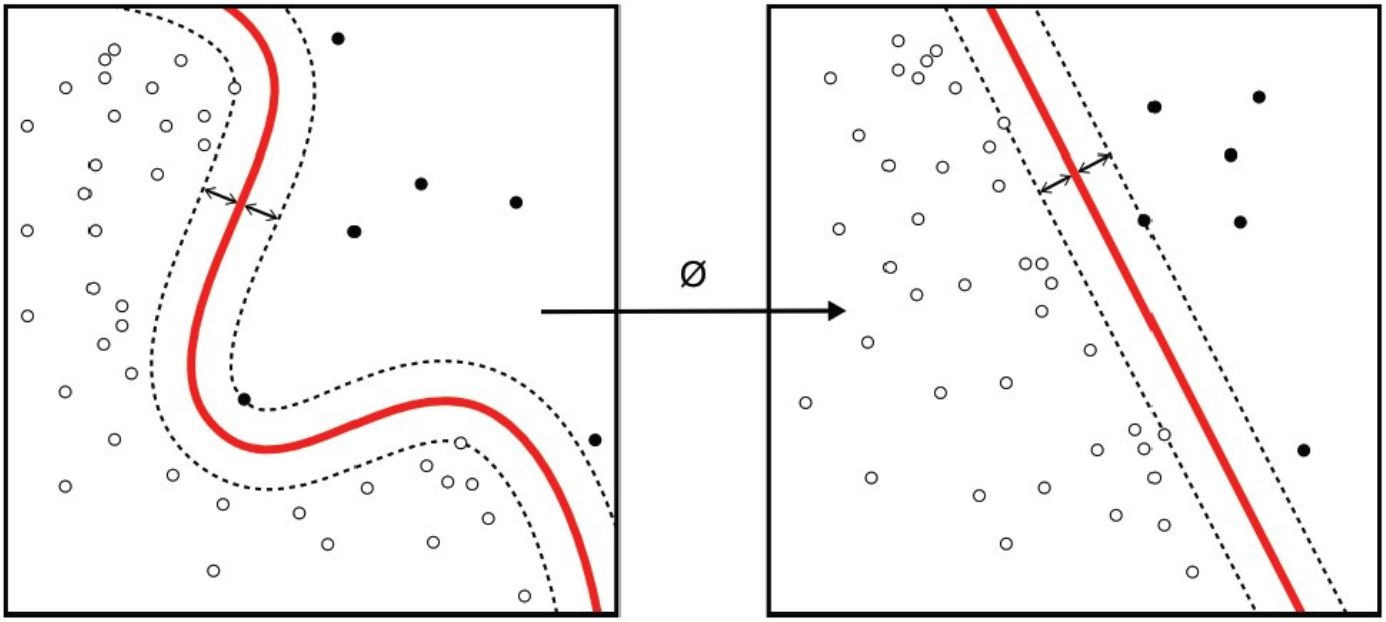
ce	stock_p_change	SP500	sp500_p_change	Difference	DE Ratio	...	Book Value Per Share	Cash Flow	Beta	Held by Insiders	Held by Institutions	Shares Short (as of	Short Ratio	Short % of Float	Shares Short (prior	Status
79	11.296044	1254.05	10.867009	0.429035	54.360	...	11.360	1.385000e+10	1.550	0.25	82.10	4360000.0	1.200	1.40	5420000.0	outperform
04	17.435198	1650.34	45.901886	-28.466688	44.120	...	15.410	1.320000e+09	1.660	0.23	83.20	4370000.0	1.200	1.30	3080000.0	underperform
68	30.095498	1655.17	46.328892	-16.233394	56.390	...	14.460	1.260000e+09	1.810	0.23	83.50	2780000.0	1.200	0.80	3300000.0	underperform
26	0.000000	1129.44	0.000000	0.000000	0.595	...	14.078	2.530000e+09	1.808	1.00	79.36	12540000.0	2.497	1.46	12700000.0	underperform
89	-3.998832	1118.38	-0.979246	-3.019586	0.561	...	14.049	1.780000e+09	1.676	1.00	76.64	7780000.0	1.935	0.90	7930000.0	underperform

GETTING CURRENT DATA

The current data can either be acquired using parsing yahoo finance or from a data provider. If you are subscribed to a data provider, they can provide you with daily OHLC prices and quarterly key statistics. If you are using web parsing to get the data, you should parse the website once every 3 months, and update your stock list. Data providers update their technical data on a daily basis, with quarterly release of fundamental data.

SUPPORT VECTOR MACHINES

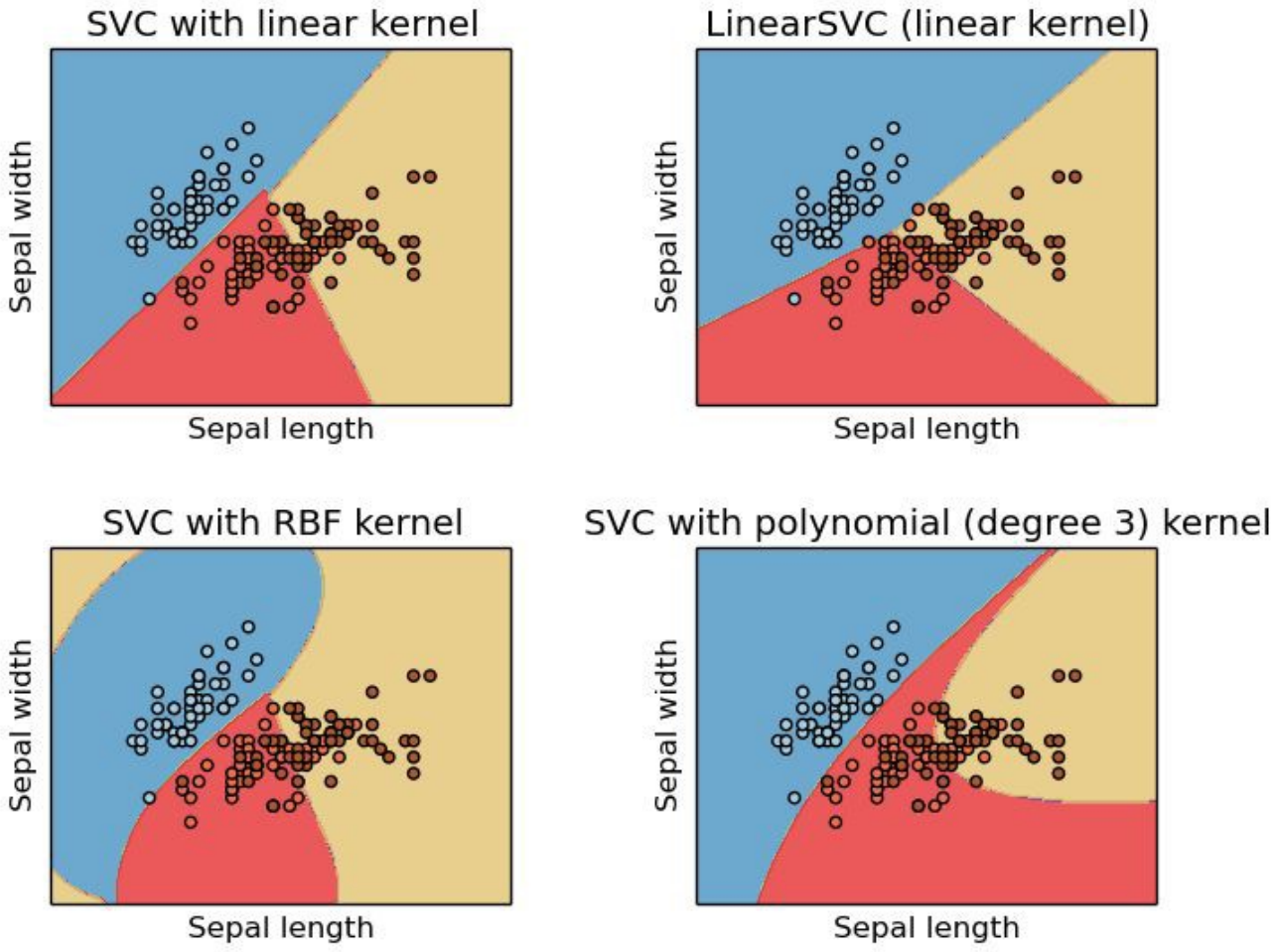
SVM is a supervised machine learning model with associated learning algorithms that are used for analyzing and predicting probabilities, using the given data. It used to perform classification and regression analysis on the given structured and unstructured data. SVM is a non-probabilistic linear classifier. Given a labeled set of training examples, each marked as belonging to one or the other. The SVM training algorithm builds a model that assigns new examples to one category or the other. It is a representation of the examples as points in space, mapped so that the data points are separated clearly. Additional new data plots are then incorporated into that same space. Based on where the point falls on, it is added to specific categories. Support Vector Machines can also perform non linear classification using the kernel trick.



SUPPORT VECTOR CLUSTERING

Clustering is to partition a data set into different groups according to some criterion in an attempt to organize data into a more meaningful form. There are many ways of achieving this form. Clustering may proceed according to some parametric model or by grouping data points according to some distance or similarity measure as in hierarchical clustering. It usually adds cluster boundaries within regions of the data space where there is insufficient data in the probability distribution area. This is the path taken in support vector clustering, which is based on the support vector approach.

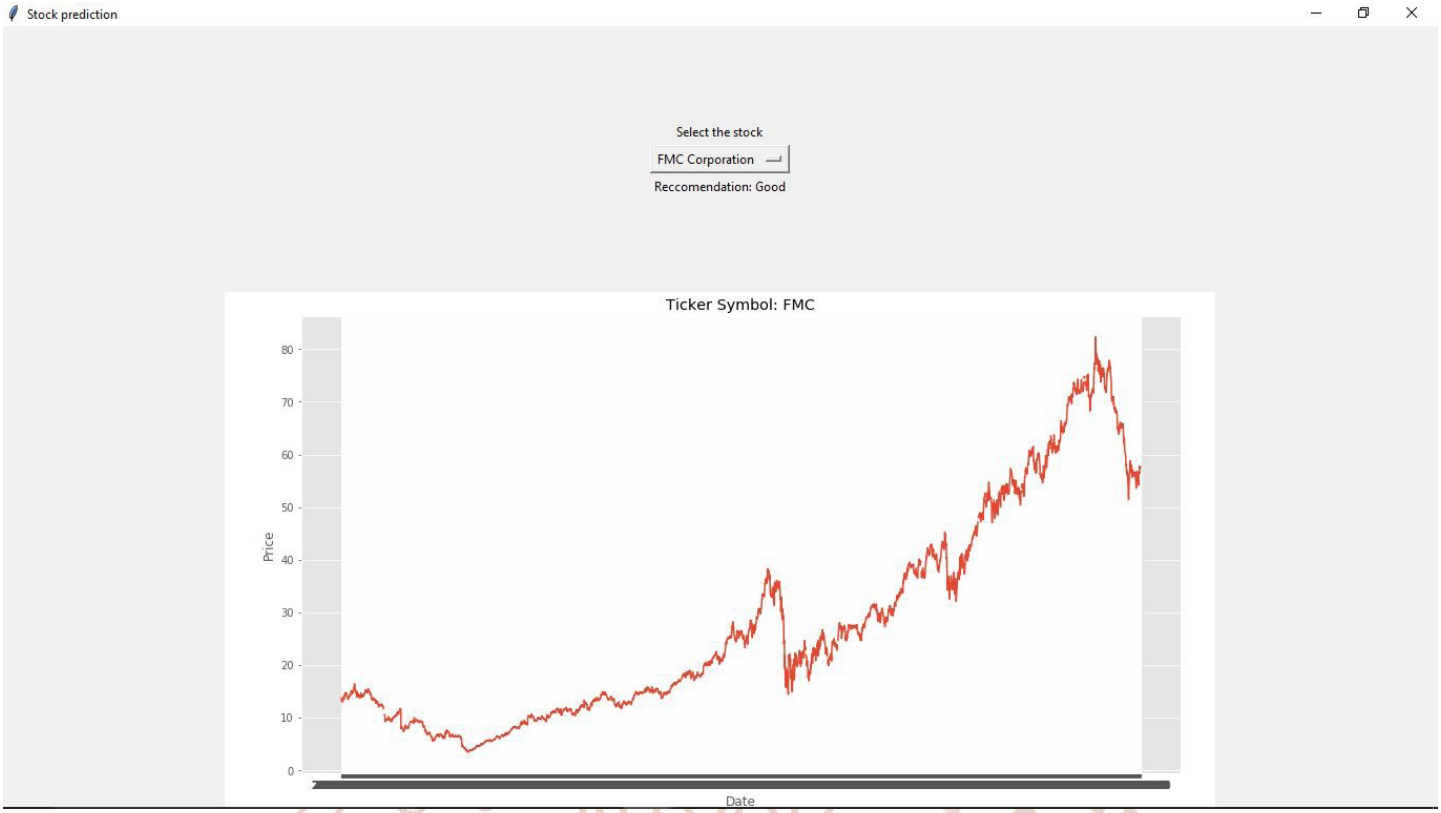
In SVC data points are mapped from data space to a high dimensional feature space using the kernel function. In feature space we look for the smallest sphere that encloses the image of the data points using the Support vector domain description algorithm (DDA). This sphere, when mapped back into the data space, will form a set of contours which can enclose the data points. We interpret these contours as cluster boundaries, and points enclosed by each contour are associated by support vector clustering to the same cluster.



PREDICTION

The multidimensional and multivariate data is then structured and passed to the prediction model. Linear Support Vector Clustering algorithm is used for predicting the future price of the stock. We can also compare the predicted stock's performance with the s&p 500 benchmark.

The accuracy of prediction and the percentage gain of the predicted stock compared to the benchmark can be interpreted.



CONCLUSION

This paper proposes a SVM-based stock market trend prediction system based on the fundamental statistics and stock prices to predict the long term movement of the stock. The system takes into account a good feature subset, which contains features that are highly correlated with the output, yet uncorrelated with each other. The selected features are evaluated carefully and prioritized. The feature selection and feature evaluation are filtered by correlation-based SVM. It reduces dimension and noise of financial data as well as provides pre selected stocks for experts and investors to make a investing decision. In the proposed system, the setting of parameters have a critical impact on the performance of the resulting system. We need to investigate to develop a structured method of selecting an optimal value for the parameters in the proposed prediction system for the best results.

REFERENCES

1. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. In Proc. of the International Conference on Computer Vision (ICCV),1999.
2. K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. The Journal of Machine Learning Research, 2:265–292, 2002.
3. K. M. Cremers and A. Petajisto. How active is your fund manager? a new measure that predicts performance. Review of Financial Studies, 22(9):3329–3365, 2009.
4. C.-M. Hsu. A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming. Expert Systems with Applications, 38 (11):14026–14036, 2011
5. S. Gould. Learning weighted lower linear envelope potentials in binary markov random fields. IEEE transactions on pattern analysis and machine intelligence, 37(7):1336–1346, 2015.
6. W. Huang, Y. Nakamori, and S.-Y. Wang, “Forecasting stock market movement direction with support vector machine,” Computers and Operations Research, vol. 32, pp. 2513–2522, 2005.
7. J. D. Piotroski, “Value investing: The use of historical financial statement information to separate winners from losers,” Journal of Accounting Research, vol. 38, pp. 1–41, 2000.