STOCK MARKET PREDICTION

PREFACE

For making this project we have studied various concepts related to the stock market and how they can be used. We also studied about various Machine Learning algorithms and tools that can be used to solve the problem easily.

The project aims at applying two machine learning algorithms; Decision Trees and Support Vector Machines and analyze how these algorithms performs at predicting the stock market.

ACKNOWLEDGEMENT

Apart from the efforts of ourselves, the success of any project depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental in the successful completion of this project. The guidance and support received from all the members who contributed and who are contributing to this project, was vital for the success of the project.

ABSTRACT

The prediction of a stock market direction may serve as an early recommendation system for short-term investors and as an early financial distress warning system for long-term shareholders. Forecasting accuracy is the most important factor in selecting any forecasting methods. Research efforts in improving the accuracy of forecasting models are increasing since the last decade. The appropriate stock selections those are suitable for investment is a very difficult task. The key factor for each investor is to earn maximum profits on their investments.

In this paper Support Vector Machine Algorithm (SVM) is used. SVM is a very specific type of learning algorithms characterized by the capacity control of the decision function, the use of the kernel functions and the scarcity of the solution. In this paper, we investigate the predictability of financial movement with SVM. To evaluate the forecasting ability of SVM, we compare its performance with Decision trees.

These methods are applied on data retrieved from Yahoo Finance. The results will be used to analyze the stock prices and their prediction in depth in future research efforts.

OBJECTIVE

In the past decades, there is an increasing interest in predicting markets among economists, policymakers, academics and market makers. The objective of the proposed work is to study and improve the supervised learning algorithms to predict the stock price.

Technical Objective

The system must be able to access a list of historical prices. It must calculate the estimated price of stock based on the historical data. It must also provide an instantaneous visualization of the market index.

Experimental Objective

Two versions of prediction system will be implemented; one using Decision trees and other using Support Vector Machines. The experimental objective will be to compare the forecasting ability of SVM with Decision Trees. We will test and evaluate both the systems with same test data to find their prediction accuracy.

STATEMENT OF THE PROBLEM

Financial analysts investing in stock market usually are not aware of the stock market behavior. They are facing the problem of trading as they do not properly understand which stocks to buy or which stocks to sell in order to get more profits. In today's world, all the information pertaining to stock market is available. Analyzing all this information individually or manually is tremendously difficult. As such, automation of the process is required. This is where Datamining techniques help.

Understanding that analysis of numerical time series gives close results, intelligent investors use machine learning techniques in predicting the stock market behavior. This will allow financial analysts to foresee the behavior of the stock that they are interested in and thus act accordingly.

The input to our system will be historical data from Yahoo Finance. Appropriate data would be applied to find the stock price trends. Hence the prediction model will notify the up or down of the stock price movement for the next trading day and investors can act upon it so as to maximize their chances of gaining a profit. The entire system would be implemented in Python language using open source libraries.

DEFINITION OF THE PROBLEM

Stock market attracts thousands of investors' hearts from all around the world. The risk and profit of it has great charm and every investor wants to book profit from that. People use various methods to predict market volatility, such as K line diagram analysis method, Point Data Diagram, Moving Average Convergence Divergence, even coin tossing, fortune telling, and so on.

THEORETICAL BACKGROUND OF THE PROBLEM

Stock market is highly volatile. At the most fundamental level, it is said that supply and demand in the market determines stock price. But, it does not follow any fixed pattern and is also affected by a large number of highly varying factors

The investors on the Wall Street are split in two largest factions of adherents; those who believe the market cannot be predicted and those who believe the market can be beat.

RELATED RESEARCH TO SOLVE THE PROBLEM

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

Recent researches uses input data from various sources and multiple forms. Some systems use historical stock data, some use financial news articles, some use expert reviews while some use a hybrid system which takes multiple inputs to predict the market.

Also, a wide range of machine learning algorithms are available that can be used to design the system. These systems have different approaches to solve the problem. Some systems perform mathematical analysis on historic data for prediction while some perform sentiment analysis on financial news articles and expert reviews for prediction.

OUR SOLUTION TO SOLVE THIS PROBLEM

We will implement the system using two different machine learning techniques. One using Support Vector Machines and the second implementation using DecisionTrees.

We will train both the systems using 75% of historic data and then test our model to check which systems yields better output using the remaining 25% of historic data.

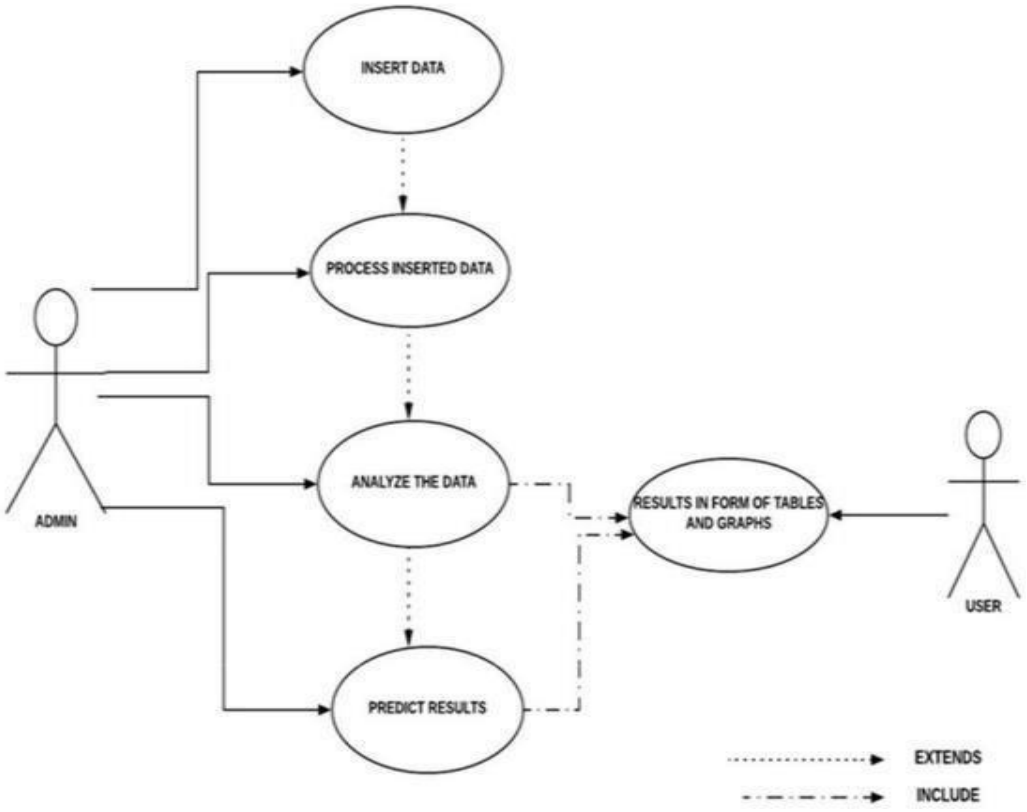
WHY OUR SOLUTION IS BETTER?

It uses SVM and Decision Trees which have better performance than Neural Network. Moreover, using SVM will takes away the burden of matching the present price pattern with historic patterns and also SVM trains faster than a NN and has a lower computational cost.

Also, other solution uses the financial data as it is without using any indicators, whereas our solution uses many indicators such as EMA, RSI, MACD, SMI, CCI, ROC, CMO, WPR and ADX to get better results.

Use case Diagram

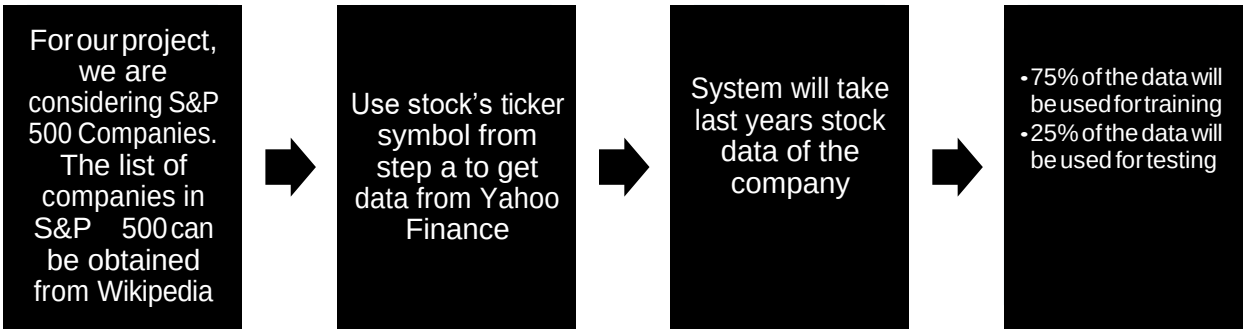
A dynamic and behavioral diagram in UML is use case diagram. Use cases are basically set of actions, services which are used by system. To visualize the functionality requirement of the system this use case diagram are used. The internal and external events or party that may influence the system are also pictured. Use case diagram specify how the system acts on any action without worrying to know about the details how that functionality is achieved.



HOW TO COLLECT INPUT DATA?

Input data is taken from Yahoo Finance using following steps:

1. For our project, we are considering S&P 500 Companies. The list of companies in S&P 500 can be obtained from Wikipedia [3].
2. Use stock's ticker symbol from step a to get data from Yahoo Finance.
3. System will take last years stock data of the company .
4. Further we divide the data into two parts, training data and testing data, where 75% of the data will be used for training and 25% of the data will be used for testing.



HOW TO SOLVE THE PROBLEM?

To solve the problem, we will follow below steps ---

5. Fetch the data of a stock from Yahoo Finance of last years.
6. Calculate the values of technical indicators RSI, EMA, MACD, SMI, etc.
7. Train the model using these indicators and training data.
8. Test the model using testing data.
9. Evaluate our system using various evaluation techniques.

Details of Technical Indicators ---

- **Relative Strength Index --- RSI**
  The relative strength index (RSI) is a technical momentum indicator that compares the magnitude of recent gains to recent losses in an attempt to determine overbought and oversold conditions of an asset. It is calculated using the following formula:

  $$RSI = 100 - 100 / (1 + RS^*)$$

  Where, RS = Average of x days' up closes / Average of x days' down closes.

- **Exponential Moving Average --- EMA**
  An exponential moving average (EMA) is a type of moving average that is similar to a simple moving average, except that more weight is given to the latest data. The exponential moving average is also known as "exponentially weighted moving average".

- **Moving Average Convergence Divergence --- MACD**
  Moving average convergence divergence (MACD) is a trend-following momentum indicator that shows the relationship between two moving averages of prices. The MACD is calculated by subtracting the 26-day exponential moving average (EMA) from the 12-day EMA. A nine-day EMA of the MACD, called the "signal line", is then plotted on top of the MACD, functioning as a trigger for buy and sell signals.

- **Stochastic Momentum Index --- SMI**
  The Stochastic oscillator is a technical momentum indicator that compares a security's closing price to its price range over a given time period. The oscillator's sensitivity to market movements can be reduced by adjusting the time period or by taking a moving average of the result. This indicator is calculated with the following formula:
  %K = 100 [(C --- L14) / (H14 --- L14)]
  Where C = the most recent closing price
  L14 = the low of the 14 previous trading sessions
  H14 = the highest price traded during the same 14-day period.
  %D = 3-period moving average of %K

- **Commodity Channel Index --- CCI**
  An oscillator used in technical analysis to help determine when an investment vehicle has been overbought and oversold. The Commodity Channel Index, first developed by Donald Lambert, quantifies the relationship between the asset's price, a moving average (MA) of the asset's price, and normal deviations (D) from that average. It is computed with the following formula:

  $$CCI = \frac{price - MA}{0.015 \times D}$$

- **Collateralized Mortgage Obligation --- CMO**
  A collateralized mortgage obligation (CMO) is a type of mortgage-backed security in which principal repayments are organized according to their maturities and into different classes based on risk. A collateralized mortgage obligation is a special purpose entity that receives the mortgage repayments and owns the mortgages it receives cash flows from (called a pool). The mortgages serve as collateral, and are organized into classes based on their risk profile. Income received from the mortgages is passed to investors based on a predetermined set of rules, and investors receive money based on the specific slice of mortgages invested in (called a tranche).

- **Rate of Change --- ROC**
  The price rate of change (ROC) is a technical indicator that measures the percentage change between the most recent price and the price "n" periods in the past. It is calculated by using the following formula:

  > (Closing Price Today --- Closing Price "n" Periods Ago) / Closing Price "n" Periods Ago

  ROC is classed as a price momentum indicator or a velocity indicator because it measures the rate of change or the strength of momentum of change.

- **Average Directional Index --- ADX**
  The average directional index (ADX) is an indicator used in technical analysis as an objective value for the strength of trend. ADX is non-directional so it will quantify a trend's strength regardless of whether it is up or down. ADX is usually plotted in a chart window along with two lines known as the DMI (Directional Movement Indicators). ADX is derived from the relationship of the DMI lines.

- **Williams % R - WPR**
  Williams %R, in technical analysis, is a momentum indicator measuring overbought and oversold levels, similar to a stochastic oscillator. It was developed by Larry Williams and compares a stock's close to the high-low range over a certain period of time, usually 14 days.

ALGORITHM DESIGN

Using Decision Trees

Step 1: Get the required data [Date, Open, High, Low, Close, Volume, Adjusted]

Step 2: Calculate all the indicator required indicator

Step 3: Calculate the prediction variable (Up/Down)

Step 4: Build the decision tree from the data calculated in above steps

Step 5: Prune the tree to remove any overfitting of data. Step 6: Give test data to the trees.

Using Support Vector Machines

We will be using C-classification Support Vector Machine with RBF Kernel. Step 1: Read the required data [Date, Open, High, Low, Close,Volume, Adjusted].

Step 2: Calculate all the required indicators ---

Step 3: Calculate the prediction variable (Up/Down)

Step 4: Provide the data from above steps to train SVM (RBF, C = 1, gamma = ½)Step 5:

Provide test data and display the results

Step 6: Compare the output from step 5 and 6 and show the observations.

HOW TO GENERATE OUTPUT?

Perform following steps to generate output:

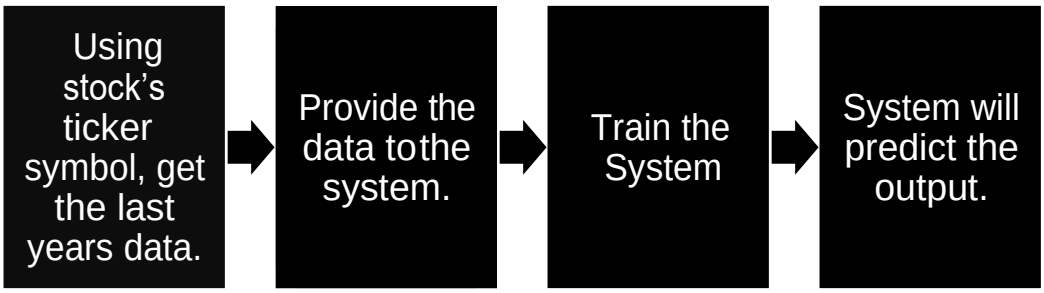
Using stock's ticker symbol, get the last years data.

Provide the data to the system.

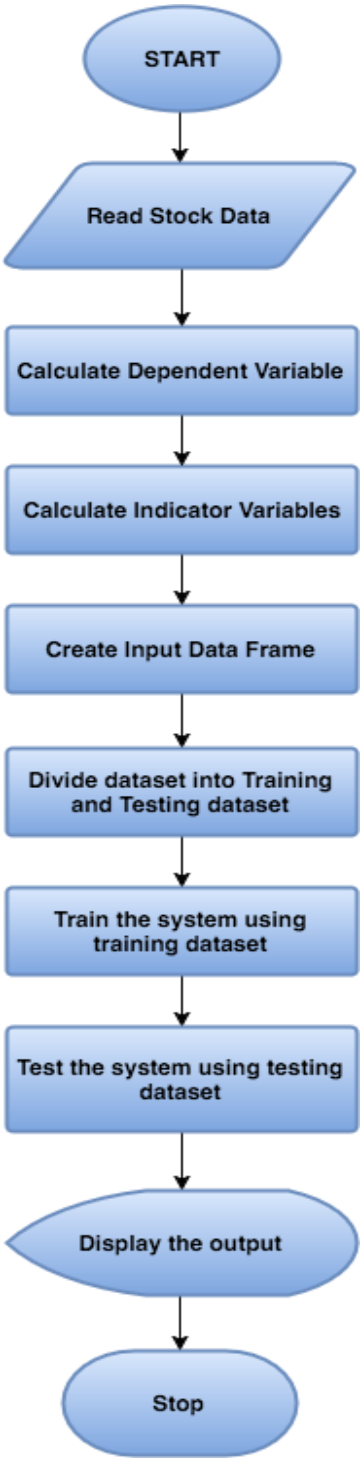
It will Train the system.

System will predict the

output.



Flowchart


START
Read Stock Data
Calculate Dependent Variable
Calculate Indicator Variables
Create Input Data Frame
Divide dataset into Training and Testing dataset
Train the system using training dataset
Test the system using testing dataset
Display the output
Stop

OUTPUT ANALYSIS

The close analysis of the output of Decision trees and SVM algorithm reveals that the SVM gives better results than decision trees. Table 2 displays a comparison of the output of both the algorithms.

Stock Name	Parameters	Decision Tree	SVM
Apple	RSI	48.81	46.45
	WPR	60.62	59.84
	ADX	55.11	50.39
	CMO	58.26	53.54
	CCI	59.05	62.99
	ROC	65.35	63.77
	EMA Cross	50.39	50.39
	Combined Accuracy	**72.44**	**80.31**
	RMSE	**0.525**	**0.444**
Microsoft	RSI	57.48	49.60
	WPR	57.48	60.62
	ADX	51.96	49.60
	CMO	60.62	57.48
	CCI	61.41	64.56
	ROC	69.29	66.92
	EMA Cross	47.24	49.60
	Combined Accuracy	**81.10**	**82.67**
	RMSE	**0.435**	**0.416**
IBM	RSI	52.75	51.96
	WPR	62.99	66.92
	ADX	51.96	51.96
	CMO	51.18	59.05
	CCI	61.41	62.99
	ROC	70.86	71.65
	EMA Cross	51.96	51.18
	Combined Accuracy	**76.37**	**85.03**
	RMSE	**0.486**	**0.387**

General Motors	RSI	53.54	53.54
	WPR	58.26	66.14
	ADX	49.60	48.81
	CMO	62.99	61.41
	CCI	59.05	66.14
	ROC	70.07	70.07
	EMA Cross	48.81	48.81
	Combined Accuracy	**74.01**	**82.67**
	RMSE	**0.509**	**0.416**
General Electric	RSI	51.96	40.94
	WPR	66.92	66.92
	ADX	40.15	39.37
	CMO	49.6	66.92
	CCI	62.20	66.92
	ROC	65.35	64.56
	EMA Cross	44.88	40.94
	Combined Accuracy	**73.22**	**82.67**
	RMSE	**0.517**	**0.416**
Facebook	RSI	58.26	48.81
	WPR	65.35	62.99
	ADX	54.33	51.18
	CMO	56.69	55.9
	CCI	63.77	65.35
	ROC	67.71	70.07
	EMA Cross	58.18	58.26
	Combined Accuracy	**78.74**	**87.4**
	RMSE	**0.461**	**0.355**
Google	RSI	47.24	48.03
	WPR	60.62	66.14
	ADX	50.39	48.81
	CMO	56.69	56.69
	CCI	64.56	63.77
	ROC	70.86	67.7
	EMA Cross	48.03	48.03
	Combined Accuracy	**80.31**	**81.88**
	RMSE	**0.444**	**0.425**

COMPARE OUTPUT AGAINSTHYPOTHESIS

The prediction accuracy depends upon the choice of indicator variables. We tried multiple indicator variables and their permutations and selected the permutation which gave best result. With the chosen combination of indicator variables, we were able to get the maximum accuracy of 87.4% for Facebook stock.

STATISTIC REGRESSION

The independent variables used are RSI, EMA Crossover, CCI, ROC, CMO, WPR and ADX. The effect of each of these independent variables on the accuracy of output can be seen in Table 2.

DISCUSSION

We selected stocks of 7 companies to train and test the system. Two years of data is downloaded from Yahoo Finance, of which 75% is used to train the system and the remaining 25% is used for testing.

Many indicator functions and their permutations were tested while training and testing the system. Of all the indicator functions tested, the ones which gave the best prediction result were selected. The system performs very well for the prediction of the selected stocks.

SUMMARY AND CONCLUSION

In this paper, we study the use of decision trees and support vector machines to predict financial movement direction. Of both these algorithms, we saw that Support Vector Machine gave us better results. SVM is a promising type of tool for financial forecasting. SVM is superior to the other individual classification methods in forecasting daily movement direction. This is a clear message for financial forecasters and traders, which can lead to a capital gain. However, each method has its own strengths and weaknesses. In this model, the principal components identified by the SVM are used along with internal and external financial factors in SVM for forecasting. We also observed that 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.