ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Stock Market Prediction for Quantitative Trading Strategies in Financial Market Using Machine Learning Ensemble Methods

Prof.S.S.Kumbhar, Varsha Dhumal

Assistant Professor, Dept. of Computer Engineering and I.T., College of Engineering, Pune. India

Dept. of Computer Engineering and I.T., College of Engineering, Pune. India

ABSTRACT: Financial time series prediction is one of the interesting and challenging tasks nowadays, due to its potential of yielding significant profits on invested money in a short period of time. Unstable nature of the securities makes it hard to predict the next day stock prices. It is important to have a significant and well-constructed set of features to elaborate stock trends. In this paper, we have proposed a Ensemble Learning Model which predicts about daily trend of stock market, whether to take long position or short position. It also helps out in loss situations, whether to continue the strategy or not. It comprises of 2-tier framework. In first tier, we extracted some technical indicators based on some raw elements like- opening price, daily high price, daily low price, closing price trading volume etc. In second tier, we applied different classification algorithms on the extracted feature set and then combined these through Ensemble methods. we have trained the model through walk forward method and predicted the movement of daily stock trend and then evaluated its performance.

KEYWORDS: prediction, Open to close, ensemble model, stock trend, classification, long position, short position etc.

I. INTRODUCTION

The stock market is one of the most important ways for companies to raise money, along with debt markets which are generally more imposing but do not trade publicly[13].The stock market is often considered the primary indicator of a country’s economic strength and development.

- **Short selling**: In short selling, the trader borrows stock (usually from his brokerage which holds its clients’ shares or its own shares on account to lend to short sellers) then sells it on the market, betting that the price will fall. The trader eventually buys back the stock, making money if the price fell in the meantime and losing money if it rose. Exiting a short position by buying back the stock is called”covering.”
- **Margin buying**: In margin buying, the trader borrows money (at interest) to buy a stock and hopes for it to rise. Most industrialized countries have observed that if the borrowing is based on collateral from other stocks the trader owns outright, it can be a maximum of a certain percentage of those other stocks’ value.
- **Stock market index**: The movements of the prices in a market or section of a market are captured in price indices called stock market indices, of which there are many, e.g., the S and P, the FTSE and the NSE indices. Such indices are usually market capitalization weighted, with the weights reflecting the contribution of the stock to the index.

II. LITERATURE SURVEY

A. Support Vector Machine-

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis[8]. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier[8]. An SVM model is a representation of the examples as points in space, mapped

DOI: 10.15680/IJIRCCE.2016. 0404011

so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

B. Decision tree-

It is a type of supervised learning algorithm that is mostly used for classification problems. surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible [1].

C. Naive Bayes-

It is a classification technique based on Bayes theorem with an assumption of independence between predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter [2]. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier would consider all of these properties to independently contribute to the probability that this fruit is an apple.

D. KNN

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry [6]. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common among its K nearest neighbors measured by a distance function.

E. Open to Close Strategy

Open to Close strategy is a one day trading strategy. In this strategy when the market opens in the morning, the position for the trading stock is taken(LONG or SHORT).At the time of closing the market the position taken in the morning is cleared out.

III. ENSEMBLE LEARNING

Ensemble methods have been called the most influential development in Data Mining and Machine Learning in the past decade. They combine multiple models into one usually more accurate than the best of its components. Ensemble can provide a critical boost to industrial challenges – from investment timing to drug discovery, and fraud detection to recommendation systems – where predictive accuracy is more vital than model interpretability. Ensembles are useful with all modeling algorithms, but this book focuses on decision trees to explain them most clearly. After describing trees and their strengths and weaknesses, the authors provide an overview of regularization – today understood to be a key reason for the superior performance of modern ensemble algorithms. IS reveals classic ensemble methods – bagging, random forests, and boosting – to be special cases of a single algorithm, thereby showing how to improve their accuracy and speed. REs are linear rule models derived from decision tree ensembles. They are the most interpretable version of ensembles, which is essential to applications such as credit scoring and fault diagnosis. Lastly, the authors explain the paradox of how ensembles achieve greater accuracy on new data despite their (apparently much greater) complexity [21].

A. **Majority Voting**: Majority voting is the machine learning ensemble algorithm in which prediction result is decided on the majority basis. If there are three classifiers of which results are(1,-1,-1), then as the output of majority voting algorithm final prediction output will be -1.

IV. PROPOSED MODEL

The proposed model is made up of 3 components shown in figure 1.

A. Preprocessing Component-

In preprocessing component, firstly we collected the raw data from the market and processed it and then extracted some technical features or indicators based on the historical stock prices and trading volume and then we finally normalized the whole features set.

B. Prediction Component-

In Prediction Component, first we built different base models on normalized data set and then combined these models through Ensemble methods and then set a norm to predict the movement of daily's stock trend such as up or down for the next trading day from the previous day.

C. Performance Component-

In Performance Component, we compute prediction accuracy to evaluate the performance of proposed and basic individual algorithms using the output parameters like Kappa, Max dd, annualized return and sharp ratio.

- **Kappa**: Kappa calculation enables efficient risk-adjusted return measurements and comparisons among a broad range of investment alternatives.
- **Max dd**:The drawdown is the measure of the decline from a historical peak in some variable (typically the cumulative profit or total open equity of a financial trading strategy).
- **Annualized return**:It is yearly rate of return which is inferred by extrapolating returns measured over periods either shorter or longer than one calendar year.
- **Sharp ratio**:The Sharpe Ratio is a measure for calculating risk-adjusted return.
- **Accuracy**:Accuracy gives the idea of how much is the correctness of the prediction of classifier.
- **Stop Loss**:Stop loss value as the name indicates is the limiting value for the loss. In our model we have used 2 percent as the stop loss limit by default. User can change it as per the requirement. If after taking the position in the strategy, direction in the prize is leading towards loss then after reaching its stop loss limit the strategy will be squared off. And thus the further loss will be limited. Stop loss value depends on many factors such as market opening prize, closing prize, OI,high prize, low prize etc.
- **Stop Gain**:As the name indicates stop gain is the limit for the gain.Stop gain value depends on many factors such as market opening prize, closing prize, OI,high prize, low prize etc.
- **Transaction cost**:When investors purchase or sell securities via a broker or other financial intermediary, the intermediary charges a commission or fee for providing this service.This is the transaction cost. Brokers can change the value for this.Depending on this the annualized return of the clients will change.
- **Slippage factor**:With regard to futures contracts as well as other financial instruments, slippage is the difference between where the computer signaled the entry and exit for a trade and where actual clients, with actual money, entered and exited the market using the computers signals.

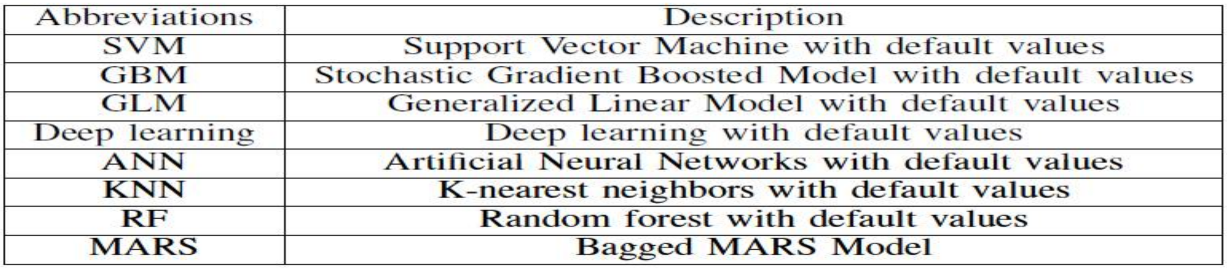
V. EXPERIMENTAL DESIGN

A. Data Set-

We collected the historical daily stock prices and trading volume from the market. We use a proprietary dataset for our experiments so not mention too much detail about it. Data set is of Nifty. It has data from around 2003 of computed and normalized 26 features.

B. Input and Output-

We have used raw data to calculate some of the technical indicators and used as a input to a model to predict the movement of daily stock trend. Input data is created by using data scrapers from various indexes including HSI, DAX S and P, NIKKIE 225, NSE NIFTY etc. The features are extracted from historical stock price and trading volume. Our model prediction for the today and the output is in the form of Binary classification. The output will decide the position for the trader whether to take long position for the stock or the short position.



Abbreviations	Description
SVM	Support Vector Machine with default values
GBM	Stochastic Gradient Boosted Model with default values
GLM	Generalized Linear Model with default values
Deep learning	Deep learning with default values
ANN	Artificial Neural Networks with default values
KNN	K-nearest neighbors with default values
RF	Random forest with default values
MARS	Bagged MARS Model

TABLE I: LIST OF ALGORITHMS

C. Walk forward Method for Training and Testing-

We used walk forward method for training and testing since underlying dynamics might change rapidly in stock markets. In Walk Forward method, we made a window of k rows, first we trained on k rows and then predicted next day stock price. It will show movement (up or down) from previous day stock price. Now, we moved the window 1 row ahead, we have taken the k rows again and trained

again. In our experiment, for walk forward method we have used variable window size and testing window of 5 rows ahead., then we executed various algorithms using various input parameters ,computed their performance at different norms.

D. Baseline Methods-

Table 1 shows the proposed and baseline methods. Majority voting and stacking are the proposed ensemble methods and others are baseline methods.

E. Evaluation Metric-

In Performance component of our proposed model, we have used accuracy,kappa,max dd, sharp ratio as metrics to evaluate the performance of our proposed and baseline methods. Accuracy is computed from confusion matrix. In confusion matrix the predictions are obtained by the column and actual class by the row of the matrix. Diagonal represents correct predictions. Prediction Accuracy is used to evaluate the classifier. Other performance metrics used are Kappa, Max DD, Annualized return, Accuracy and Sharp ratio.

F. **Conclusion-**

This paper presents a survey that uses several approaches for prediction. But in my opinion Stacking approach is the best and can be considered as an alternative to traditional individual prediction classifiers.

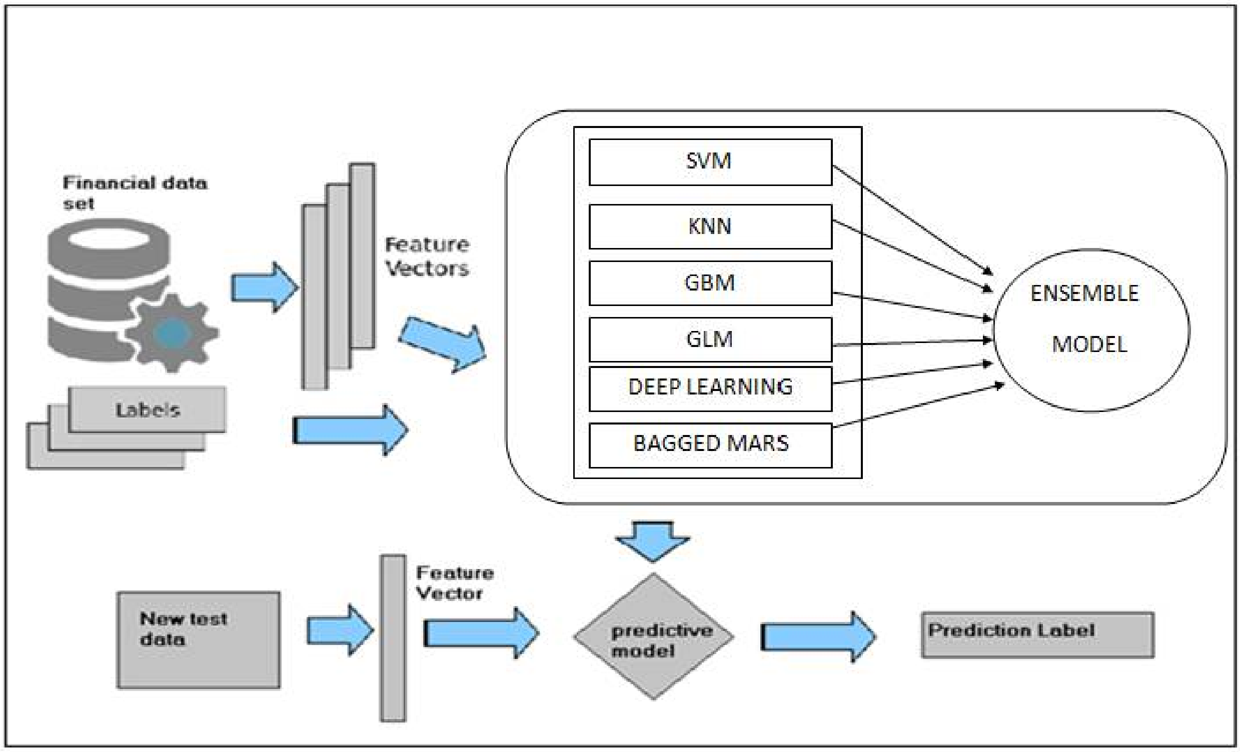


Fig. 1: Ensemble Model for Classification

REFERENCES

[1] Kuhn Johnson Applied Predictive Modeling
[2] Tom Mitchel Machine Learning
[3] Zhi Hua Zhou. Ensemble methods foundations and algorithms
[4] Machine Learning videos by Andrew Ng

[5] Rafael Thomazi Gonzalez, Prof.Dr. dante Augusto Couto, MSc.Carlos. Ensemble Systembased on Genetic algorithm for stock market forecasting Proceedings of the 978-1-4799- 7492-4/15@2015 IEEE

[6] Salim Lahmiri. ”Ensemble with radial basisFunction NeuralNetworks for Casablanca Stock Market Returns rediction”Proceedings of the-1-4799-4647-1/14@2014 IEEE

[7] Ash Booth, Enrico Gerding, Frank0.Predicting Equity Market Price Impact with Performance Weighted Ensembles of Random Forests Proceedings of the @2014 IEEE

[8] Zhen Hu 1, Jie Zhu 2, and Ken Tse 3.Stocks Market Prediction Using Support Vector Machine. International Conference on Information Management, Innovation Management and Industrial Engineering@2013

[9] http://iknowfirst.com/stock market forecast chaos theory revealing how the stock market works

[10] Dr Ashok kumar, S.Muragana. Performance Analysis of Indian Stock Market Index using Neural Network Time Series Model. Proceedings of the 2013 International Conference onPattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22

[11] Honghoi Yu, Haifei Liu.Improved Stock Market Prediction by Combining Support Vector Machine and Empirical Mode Decomposition. 2012 Fifth International Symposium on Computational Intelligence and Design

[12] Xue-ling liang, wing w. y. ng. stock investment decision support using an ensemble of l-gem based on rbfnn diverse trained from different years Proceedings of the 2012 International Conference on Machine Learning and Cybernetics, Xian, 15-17 July, 2012

[13] Cheng Cheng, Wei Xu, Jiajia Wang.A Comparison of Ensemble Methods in Financial Market Prediction Proceedings of the 2012 Fifth International Joint Conference on Computational Sciences and Optimization

[14] ”Equity market ¿ Size relative to bond markets and bank assets”.eurocapitalmarkets.org. Retrieved August 14, 2015.

[15] esari, Amedeo De; Espenlaub, Susanne; Khurshed, Arif; Simkovic, Michael (2010). ”The Effects of Ownership and Stock Liquidity on the Timing of Repurchase Transactions”.Paolo Baffi Centre Research Paper No. 2011-100. SSRN 1884171

[16] Jump upSimkovic, Michael (2009). ”The Effect of Enhanced Disclosure on Open Market Stock Repurchases”. Berkeley Business Law Journal 6 (1). SSRN 1117303.