Classification of Stock Index movement using k-Nearest Neighbours (k-NN) algorithm

M.V.SUBHA
Associate Professor – Directorate of Online and Distance Education,
Anna University of Technology, Coimbatore,
Jyothipuram, Coimbatore – 641 047, Tamil Nadu.
INDIA
subhamv@gmail.com

S.THIRUPPARKADAL NAMBI
Associate Professor – Department of Management Studies,
Guruvayurappan Institute of Management,
Palakkad Main Road, Navakkarai, Coimbatore – 641 105, Tamil Nadu,
INDIA
nambist@gmail.com

Abstract: Many research studies are undertaken to predict the stock price values, but not many aim at estimating the predictability of the direction of stock market movement. The advent of modern data mining tools and sophisticated database technologies has enabled researchers to handle the huge amount of data generated by the dynamic stock market with ease. In this paper, the predictability of stock index movement of the popular Indian Stock Market indices BSE-SENSEX and NSE-NIFTY are investigated with the data mining tool of k-Nearest Neighbours algorithm (k-NN) by forecasting the daily movement of the indices. To evaluate the efficiency of the classification technique, the performance of k-NN algorithm is compared with that of the Logistic Regression model. The analysis is applied to the BSE-SENSEX and NSE-NIFTY for the period from January 2006 to May 2011.

Keywords— Classification, Data Mining, k-Nearest Neighbours, Logistic Regression, Prediction, Stock Index movement.

1 Introduction

Modern Finance aims at identifying efficient ways to understand and visualise stock market data into useful information that will aid better investing decisions. The huge data generated by the stock market has enabled researchers to develop modern tools by employing sophisticated technologies for stock market analysis. Financial tasks are highly complicated and multifaceted; they are often stochastic, dynamic, nonlinear, time-varying, flexible structured and are affected by many economical, political factors (*Tan T.Z., C. Quek, and G.S. NG, 2007*). One of the most important problems in the modern finance is finding efficient ways of summarizing and visualizing the stock market data that would allow one to obtain useful information about the behaviour of the market (*Boginski V, Butenko S, Pardalos PM, 2005*). Although there exists numerous articles addressing the predictability of stock market return as well as the pricing of stock index financial instruments, most of the existing models rely on accurate forecasting of the level (i.e. value) of the underlying stock index or its return *(M.T.Leung et al, 2006).* Stock market prediction is regarded as the challenging task of financial time series prediction *(Kim, 2003*). Investors trade with stock index related instruments to hedge risk and enjoy the benefit or arbitrage. Therefore, being able to accurately forecast stock market index has profound implication and significance to researchers and practitioners alike *(Leung et al, 2000).*

1.1 Predictability of stock prices

There is a large body of research carried out suggesting the predictability of St ock markets. Lo and Maculay(1988) in their research paper claim that stock prices do not follow random walks and suggested considerable evidence towards predictability of stock prices. Basu (1977), Fama & French (1992), Lakonishok, Schleifer & Vishney (1997) in their various studies have carried many cross sectional analysis across the globe and tried to establish the predictability of the stock prices. Studies have tried to establish that various factors like firm size, book to market equity, and macroeconomic variables like short term interest rates, inflation, yield from short and long term bonds, GNP help in the predictability of stock returns (Fama & French (1993), Campbell (1987), Chen, Roll and Ross (1986), Cochrane (1988)). Ferson & Harvey (1991) show that predictability in stock returns are not necessarily due to market inefficiency or over-reaction from irrational investors but rather due to predictability in some aggregate variables that are part of the information set. O'Connor et al (1997) demonstrated the usefulness of forecasting the direction of change in the price level, that is, the importance of being able to classify the future return as a gain or a loss. Hence, at any given point of time, research on predictability of stock indices is significant owing to the dynamic nature of the stock markets.

1.2 Data Mining and Classification

As the dynamic stock market leaves a trail of huge amount of data, storing and analyzing these tera bytes of information has always been a challenging task for the researchers. After the advent of computers with brute processing power, storage technologies such as databases and data warehouses and the modern data mining algorithms, Information system plays a pivotal role in the stock market analysis. Data mining is the non trivial extraction of implicit, previously unknown, and potentially useful information from the data and thus it is emerging as an invaluable knowledge discovery process. The four major approaches of data mining are classification, clustering, association rule mining and estimation.

Classification is process of identifying a new objects or events as belonging to one of the known predefined classes. A marketing campaign may receive customer response or not, a customer can be a gold customer or default customer, a credit card transaction can be a genuine one or a fraudulent one, stock market can be bullish or bearish. Business manager's job is to evaluate those events and to take appropriate decisions and for that he needs to rightly classify the events.

Any object or an event can be described by a set of attributes that constitute independent variables and the outcome – which is the dependent variable- is to classify this event or an object into a set of predefined classes like 'good or bad', 'success or failure' etc. So the aim of the classification is to construct a model by discovering the influence of independent variables on the dependant variable so that unclassified records can be classified. Classification is widely used in the fields of bioinformatics, medical diagnosis, fraud detection, loan risk prediction, financial market position etc. This study employed the logistic regression model which is a very popular statistical classifier and the data mining classifier k-Nearest Neighbours (k-NN) to classify the index movement of the popular Indian stock market indices BSE-SENSEX and NSE-NIFTY.

The remaining part of this paper is organized as follows. Section 2 reviews relevant literature. Section 3 describes the source of the secondary data used for the study, its period, split of the training dataset and the test dataset etc. Section 4 explains the methodology employed in this study. Section 5 presents the results and discussion. Section 6 concludes this paper.

2 Review of Past Studies

2.1 Classical Statistical methods

Logistic regression (Hosmer & Lemeshow 1989; Press & Wilson, 1978; Studenmund, 1992, Kumar & Thenmozhi, 2006) and Multiple Regression (Menard, 1993; Myers, 1990; Neter, Wasserman & Katner 1985; Snedicor & Cochran 1980) are the classical statistical methods that have been widely employed in various studies relating to stock market analysis. Statistics has been applied for predicting the behavior of the stock market for more than half a century and a hit rate of 54% is considered as a satisfying result for stock prediction. To reach better hit rates, other data mining techniques such as neural networks are applied for stock prediction. The most obvious advantage of the data mining techniques is that they can outperform the classical statistical methods with 5–20% higher accuracy rate.(N.Ren M.Zargham & S. Rahimi 2006)

2.2 Data mining classification models

Quahand Srinivasan (1999) proposed an Artificial Neural Network (ANN) stock selection system to select stocks that are top performers from the market and to avoid selecting under performers. Huarng and Yu (2005) proposed a Type-2 fuzzy time series model and applied it to predict TAIEX. Boginski V, Butenko S, Pardalos PM (2005), in their study consider a network representation of the stock market data referred to as the market graph, which is constructed by calculating cross-correlations between pairs of stocks and studied the evolution of the structural properties of the market graph over time and draw conclusions regarding the dynamics of the stock market development. Kim (2006) proposed a genetic algorithm (GA) approach to instance selection in artificial neural networks for financial data mining. Instances are a collection of training examples in supervised learning and instance selection chooses a part of the data that is representative and relevant to the characteristics of all the data. Muh-Cherng et al. (2006) present a stock trading method by combining the filter rule and the decision tree technique. The results show that the new method outperforms both the filter rule and the previous method. Liao, Javier et al (2009), in their study adapts the k-Nearest Neighbours (k-NN) algorithm to forecast HTS and, more generally, to deal with histogram data. The proposed k-NN relies on the choice of a distance that is used to measure dissimilarities between sequences of histograms and to compute the forecasts. Chih-Fong Tsai et al, (2010) in their study used a combination of multiple feature selection methods to identify more representative variables for better stock prediction. In many studies we find the use of many popular statistical tools, in particular, three well-known feature selection methods - Principal Component Analysis (PCA), Genetic Algorithms (GA) and decision trees (CART) are used.

The present study attempts to apply the k-NN algorithm for the task of prediction and classification of stock index movement. To evaluate the efficiency of the prediction and classification technique the performance of k-NN is compared with the Logistic Regression model. This model is applied to the two popular Indian stock market indices, the Bombay Stock Exchange Sensitivity Index(BSE-SENSEX) and the NSE-NIFTY in order to classify the stock index trend that would indicate the investors whether the stock indices are likely to increase (bull) or fall (bear) in the forthcoming days. Thus the main objective of this paper is to classify the predicted stock index into bullish or bearish states.

3 Data and Sources of Data

This paper examines the daily change of closing values of BSE-SENSEX and NSE-NIFTY based on the following predictors: Open price, High price, Low price and Close price. BSE-SENSEX and NSE-NIFTY values are obtained from the BSE and NSE websites respectively for the period from Jan' 2006 to May 2011 with a sample of 1341 trading days. The data is divided into two sub-samples in the split up of 80:20 where the in-sample or training data spans from Jan' 2006 to June' 2010 with 1110 trading days and the data for the remaining period from July 2010 to May 2011 with 231 trading days are used for out-of sample or test data.

4 Methodology

4.1. k-NN algorithm

The *k*-nearest neighbours algorithm is one of the simplest machine learning algorithms. It is simply based on the idea that "objects that are 'near' each other will also have similar characteristics. Thus if you know the characteristic features of one of the objects, you can also predict it for its nearest neighbour." k-NN is an improvisation over the nearest neighbour technique. It is based on the idea that any new instance can be classified by the majority vote of its 'k' neighbours, - where *k* is a positive integer, usually a small number.

k-NN algorithm looks for 'k' nearest records with in the training dataset and uses the majority of the classes of the identified neighbours for classifying a new record. To do that, first of all, k nearest neighbours of a new instance has to be identified. The nearness is measured in terms of the distance between the new unclassified instance and old classified instances in the training dataset. One of the most widely used metric is the ***Euclidean distance.*** The Euclidean distance between two instances (x_1, x_2, x_3,…x_p) and (u_1, u_2,u_3,… u_p) is given by the following formula

$$\sqrt{(x_1-u_1)^2+(x_2-u_2)^2+...+(x_p-u_p)^2} \quad \text{-- (1)}$$

where, x_1, x_2, x_3,x_p are predictors of the instance #1 and u_1, u_2,u_3,… u_p are predictors of the instance #2.

4.1.1 k-NN prediction

The prediction model considers Opening value, High value, Low value and Closing value of the market index as independent variables and the next day's closing value as the dependent variable. The k-NN algorithm identifies 'k' nearest neighbours in the training data set in terms of the Euclidean distance with respect to the day for which prediction is to be done. Once k-nearest neighbours are identified, the prediction for that day is computed as the average of the next day's closing prices of those neighbours. The k-NN algorithm tries out various 'k' values in the training data set and finds the optimum value of k that produces the best prediction result. Then this predictive model, with the optimum value of 'k', is applied on the test data set for predicting the next day's closing value. The output of the predictive model is compared with the actual values of the test dataset for validation.

4.1.2 k-NN Classifier

The classifier model considers opening value, high value, low value, closing value and returns of the market index as independent variables and the next day's class as the dependent variable. Returns for a day is calculated as

$$returns = \frac{(v_t - v_{t-1})}{v_{t-1}} \quad \text{-- (2)}$$

Where v_t is the closing value of the index on the current day and v_{t-1} is the closing value of the index of previous day. If the next days' return is positive, the next day's class is classified as "bull" otherwise "bear". The k-NN algorithm identifies 'k' nearest neighbours in the training data set in terms of the Euclidean distance with respect to the day for which price index movement is to be classified. Once k-nearest neighbours are identified, the classification for that day is simply the majority of the next day's classes of those identified neighbours. The k-NN algorithms tries out various k values in the training dataset and finds the optimum value of k that produces the best result. Then this classifier model with the optimum value of 'k' is applied on the test data set for classifying the next day's class. In other words, it simply tries to predict the index movement for the next day either as 'bull' or "bear". The output of the classifier is compared with the actual classes of the test data set for validation.

4.2 Logistic Regression

Multiple regression is a predictive statistical technique that requires all the independent variables to be continuous and the dependent variable is also numeric. Logistic regression is the extension of this idea where the dependent variable is categorical and the independent variables can be continuous, categorical or even a combination of the two. The goal of the logistic regression is to classify the new event into one of the predefined classes. Thus it is a statistical classifier. The regression equation is of the form

$$Y= \beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots \text{ --(3)}$$

Where y is the dependent variable; x_1,x_2,x_3 are independent variables; $\beta_0,\beta_1,\beta_2,\beta_3$ are the y intercept and regression coefficients respectively. To convert the Y value into probability value, we compute logistic response function

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_3 x_3)} \quad \text{--(4)}$$

By appropriately choosing a cut-off probability value, y can be classified into any of the predefined classes. This model considers daily Opening value, High value, Low vale and Closing value of the market index as independent variables and the next day's closing price as the dependent variable which is categorical – either bull or bear. With the help of the regression equation obtained from the training data set, a logistic response function is formulated and that is used to calculate the probability of next day being bull for all the days in the test data set. With the cut-off probability value of 0.5, each day is classified into a bull state if the computed probability is higher than the cut off probability; otherwise it is classified as bear state.

5 Results and Discussion

5.1 Prediction for BSE-SENSEX

The result of the actual movement of BSE-SESEX for the test dataset of 231 trading days for the period from July 2010 to May 2011 is compared with the k-NN prediction for the same period. The results are as shown below:

Fig.1 - BSE-SENSEX ACTUAL MOVEMENT

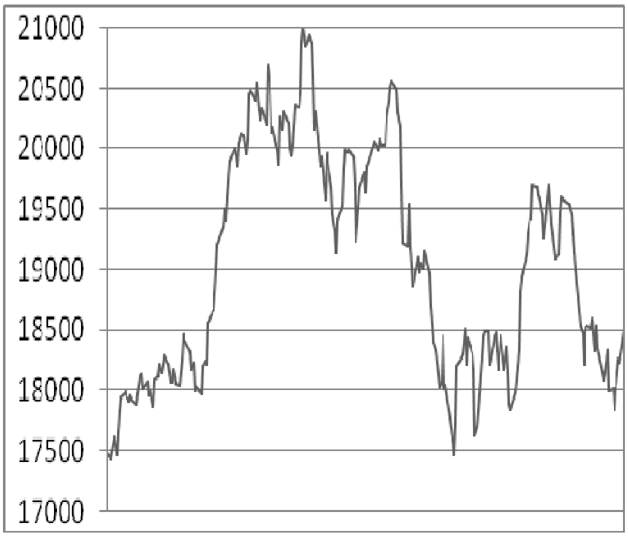
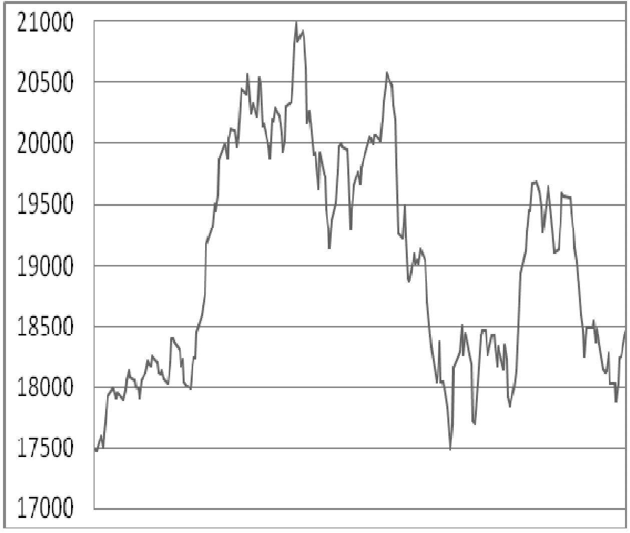


Figure -1 shows the actual movement of BSE-SENSEX values for the period from July 2010 to May 2011.

Fig.2 - BSE-SENSEX PREDICTED MOVEMENT



The above figure shows the line chart of the predicted values of the k-NN model for the same period of July 2010 to May 2011. It is seen from the figure-1 and figure-2, that the result of the k-NN predictive model very closely follows the actual movement of BSE-SENSEX.

Fig.3 – SCATTER PLOT OF RESIDUALS

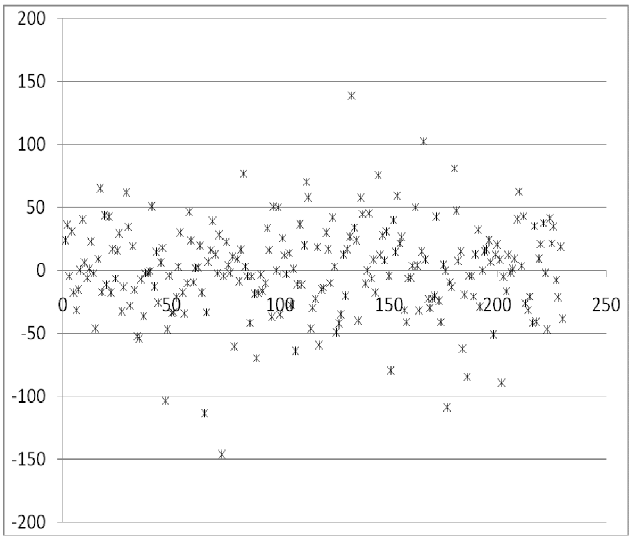


Figure-3 is the scatter plot of the *residuals*. Residuals are the measure of forecasting error, which is computed by subtracting the actual value from the forecasted value. It is observed from the above figure that most of the residuals falls under the range of -50 to +50 and only very few outliers are exceeding this range. The forecasting error of the predictive model is tabulated in the following table.

Table 1
RESULT OF K-NN PREDICTIVE MODEL

Evaluation on test set for BSE-SENSEX (k=5)	
Correlation coefficient	0.9992
Mean absolute error	27.0719
Root mean squared error	36.4836
Relative absolute error	3.23%
Root Relative squared error	3.91%

The correlation coefficient between the actual SENSEX values and the predicted values of the model is found to be 0.9992 implying a very high level of dependency of the predictive model. The forecasting error of the model is very low as indicated by the root mean squared error (36.48) and the mean absolute error (27.07). The prediction model also displays a very low error rate of prediction as measured by relative absolute error of 3.23% and the root relative squared error of 3.91%.

5.2 Prediction for NSE-NIFTY

The result of the actual movement of NSE-NIFTY for the test dataset of 231 trading days for the period from July 2010 to May 2011 is compared with the k-NN prediction for the same period. The results are as shown below.

Fig. 4 - NSE NIFTY ACTUAL MOVEMENT

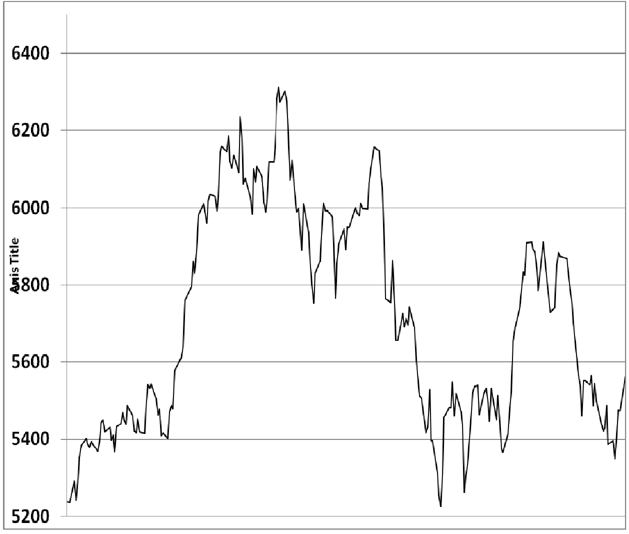
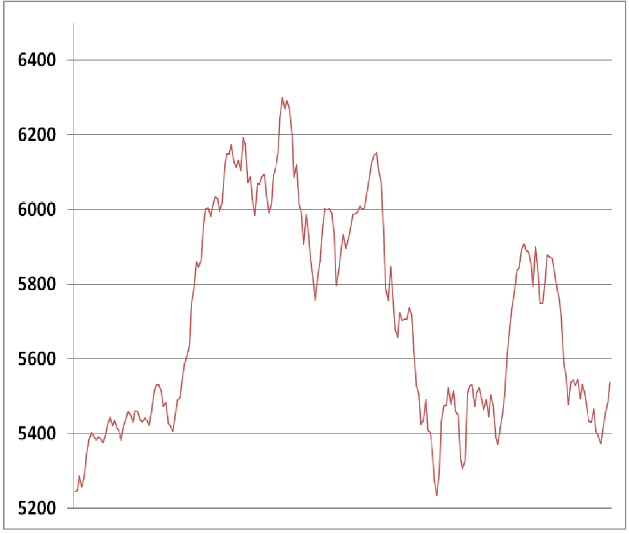


Figure -4 shows the actual movement of NSE-NIFTY values for the period from July 2010 to May 2011.

Fig. 5 - NSE NIFTY PREDICTED MOVEMENT



The above figure shows the line chart of the predicted values of the k-NN model for the same period of July 2010 to May 2011. It is seen from the figure-4 and figure-5, that the result of the k-NN predictive model very closely follows the actual movement of NSE-NIFTY.

Fig. 6 - SCATTER PLOT OF RESIDUALS

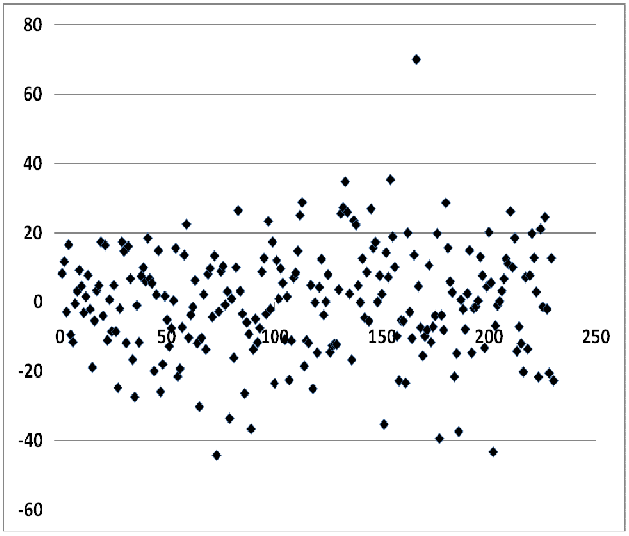


Figure-6 is the scatter plot of the *residuals*. It is observed from the above figure that most of the residuals falls under the range of -20 to +20 and only very few outliers are exceeding this range. The forecasting error of the predictive model is tabulated in the following table.

Table 2
RESULT OF K-NN PREDICTIVE MODEL

Evaluation on test set for NSE-NIFTY (k=5)	
Correlation coefficient	0.9985
Mean absolute error	12.1258
Root mean squared error	15.625
Relative absolute error	4.76 %
Root Relative squared error	5.51%

The correlation coefficient between the actual NIFTY values and the predicted values of the model is found to be 0.9985 implying a very high level of dependency of the predictive model. The forecasting error of the model is very low as indicated by the root mean squared error (12.13) and the mean absolute error (15.65). The prediction model also displays a very low error rate of prediction as measured by relative absolute error of 4.76% and the root relative squared error of 5.51%.

5.3 Classification for BSE-SENSEX

The results obtained from the two classifying models for BSE-SENSEX are given below.

Table 3

COMPARISION OF CLASSIFIER MODELS ON THE TEST DATASET FOR BSE-SENSEX

	k-NN algorithm		Logistic Regression	
Instances(231)		Accuracy		Accuracy
Correctly classified instances	205	88.74%	136	58.87%
Incorrectly classified instances	26	11.26%	95	41.13%
Kappa Statistics	0.7749		0.1777	

Table-3 shows that the k-NN algorithm rightly classifies the next day's index movement for 205 instances out of the total of 231 instances with an accuracy rate of 88.74% and misclassifies only 26 instances with an error rate of 11.26%. But the logistic regression rightly classifies the next day's index movement only for 136 instances out of the total of 231 instances with an accuracy rate of 58.87% and misclassifies 95 instances with an error rate of 41.13%. From the Kappa statistics, which is a chance corrected measure that can range from 1 to -1 indicating perfect agreement and perfect disagreement of the model respectively, it is inferred that the k-NN models shows a high degree of acceptance (0.77) compared to Logistic regression model (0.18).

Table 4

CONFUSION MATRICES FOR BSE-SENSEX

	k-NN Classifier		Logistic Regression	
	Predicted Class		Predicted Class	
	Bull	Bear	Bull	Bear
Actual Class				
Bull	100	15	71	44
Bear	11	105	51	65

Table-4 is the *confusion matrix*, which is a widely used model evaluation technique for a classifier model; where the rows represent the actual class and the column represent the predicted class. It is also known as *coincidence matrix.*

It is seen from the above table that k-NN algorithm rightly classifies 100 bull class instances out of the total of 115 bull class instances and rightly classifies 105 bear class instances out of the total of 116 bear class instances. But the logistic regression classifier rightly classifies only 71 bull class instances out of the total of 115 bull class instances and rightly classifies 65 bear class instances out of the total of 116 bear class instances.

Table 5

DETAILED ERROR REPORT OF THE CLASSIFIERS FOR BSE-SENSEX

		k-NN Classifier		Logistic Regression	
Class	# case	# error	% error	# error	% error
Bull	115	15	13.04%	44	38.26%
Bear	116	11	9.48%	51	43.97%
overall	231	26	11.26%	95	41/13%

Table-5 tabulates the class-wise predictive errors of the models. k-NN algorithm misclassifies 15 bull class instances out of the total of 115 bull class instances with an error rate of 13.04% where as the logistic regression misclassifies 44 bull class instances with an error rate of 38.26%.

For bear-instance classification, k-NN misclassifies 11 bear class instances out of the total of 116 bear class instances with an error rate of 9.48%, where as the logistic regression misclassifies 95 bear class instances out of 116 bear class instances with an error rate of 43.97%.

Table 6
DETAILED ACCURACY BY CLASS – K-NN CLASSIFIER FOR BSE-SENSEX

Class	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Bear	0.905	0.13	0.875	0.89	0.906
Bull	0.87	0.095	0.901	0.885	0.906

Table-6 lists out various model evaluation parameters such as True Positive rate, False positive rate, Precision, F-measure and Receiver operative Curve (ROC) area of both the classes for the k-NN classifier model.

Table 7
DETAILED ACCURACY BY CLASS – LOGISTIC REGRESSION FOR BSE-SENSEX

Class	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Bear	0.560	0.382	0.606	0.601	0.813
Bull	0.617	0.439	0.417	0.424	0.813

Table-7 lists out all the model evaluation parameters of both the classes for logistic regression model. It is evident from the table-6 and table-7, that k-NN algorithm outperforms the traditional logistic regression in the model evaluating parameters such as TP rate, FP Rate, Precision, F-measure and ROC area etc.

5.4 Classification for NSE-NIFTY

The results obtained from the two classifying models for NSE-NIFTY are given below.

Table 8

COMPARISION OF CLASSIFIER MODELS ON THE TEST DATASET FOR NSE-NIFTY

	k-NN algorithm		Logistic Regression	
Instances(231)		Accuracy		Accuracy
Correctly classified instances	184	79.65%	125	54.11%
Incorrectly classified instances	47	20.35%	106	45.89%
Kappa Statistics	0.6975		0.1633	

Table-8 shows that the k-NN algorithm rightly classifies the next day's index movement for 184 instances out of the total of 231 instances with an accuracy rate of 79.65% and misclassifies 47 instances with an error rate of 20.35%. But the logistic regression rightly classifies the next day's index movement only for 125 instances out of the total of 231 instances with an accuracy rate of 54.11% and misclassifies 106 instances with an error rate of 45.89%. From the Kappa statistics, it is inferred that the k-NN models shows a high degree of acceptance (0.69) compared to Logistic regression model (0.16).

Table 9

CONFUSION MATRICES FOR NSE-NIFTY

	k-NN Classifier		Logistic Regression	
	Predicted Class		Predicted Class	
	Bull	Bear	Bull	Bear
Actual Class				
Bull	83	30	63	50
Bear	17	101	56	62

It is seen from the above table that k-NN algorithm rightly classifies 83 bull class instances out of the total of 113 bull class instances and rightly classifies 101 bear class instances out of the total of 118 bear class instances. But the logistic regression classifier rightly classifies only 63 bull class instances out of the total of 113 bull class instances and rightly classifies 62 bear class instances out of the total of 118 bear class instances.

Table 10

DETAILED ERROR REPORT OF THE CLASSIFIERS FOR NSE-NIFTY

		k-NN Classifier		Logistic Regression	
Class	#case	#error	% error	# error	% error
Bull	113	30	26.55%	50	44.28%
Bear	118	17	14.40%	56	47.46%
overall	231	47	20.35%	106	45.89%

Table-10 tabulates the class-wise predictive errors of the models. k-NN algorithm misclassifies 30 bull class instances out of the total of 113 bull class instances with an error rate of 26.55% where as the logistic regression misclassifies 50 bull class instances with an error rate of 44.28%. For bear-instance classification, k-NN misclassifies 17 bear class instances out of the total of 118 bear class instances with an error rate of just 14.40%, where as the logistic regression misclassifies 56 bear class instances out of 118 bear class instances with an error rate of 47.46%.

Table 11
DETAILED ACCURACY BY CLASS – K-NN CLASSIFIER FOR NSE-NIFTY

Class	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Bear	0.856	0.265	0.702	0.742	0.877
Bull	0.735	0.144	0.747	0.698	0.877

Table-11 lists out various model evaluation parameters such as True Positive rate, False positive rate, Precision, F-measure and Receiver operative Curve (ROC) area of both the classes for the k-NN classifier model.

Table 12
DETAILED ACCURACY BY CLASS – LOGISTIC REGRESSION FOR NSE-NIFTY

Class	TP Rate	FP Rate	Precision	F-Measure	ROC Area
Bear	0.525	0.442	0.514	0.546	0.592
Bull	0.558	0.475	0.543	0.55	0.592

Table-7 lists out all the model evaluation parameters of both the classes for logistic regression model. Thus the k-NN algorithm outperforms the traditional logistic regression in the model evaluating parameters such as Accuracy rate%, Error rate%, Kappa statistics, True positive rate, False positive rate, Precision, F-Measure and area of ROC curve both in the cases of BSE-SENSEX and NSE-NIFTY.

6 Conclusion

Financial markets are highly volatile and generate huge amount of data on a day to day basis. The present study applied the popular data mining tool of k-NN for the task of prediction and classification of the stock index values of BSE-SENSEX and NSE-NIFTY. The results of k-NN forecasting model is very encouraging as the forecasting errors such as the root mean squared errors and relative absolutes errors are very small for both BSE-SENSEX and NSE-NIFTY. Also, the results of k-NN classifier are compared with the Logistic regression model and it is observed that the k-NN classifier outperforms the traditional logistic regression method as it classifies the future movement of the BSE-SENSEX and NSE-NIFTY more accurately. The outcome of this research work also shows that the k-NN classifier outperforms the traditional logistic regression method in all the model evaluation parameters such as kappa statistics, precision, % error, TPF, TNR, F-measure, ROC area etc. While there are controversial theories regarding the predictability of stock markets, this study sheds a positive light on the predictability of the stock index movements through modern data mining tools. Further studies are recommended in the area of data mining applications in stock markets to gain more useful insights about the predictability of stock markets.

References

[1] Alex Berson, Stephan Smith, Kurt Thearling (2000) *Building data Mining Applications for CRM* Tata McGraw Hill, New Delhi.

[2] Basu, S. (1977). The investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *Journal of Finance* 32, 663–682.

[3] Boginski V, Butenko S, Pardalos PM (2005), Mining market data: A network approach, *Computers & Operations Research* 33 (2006) 3171–3184.

[4] Campbell, J. (1987). Stock returns and the term structure. *Journal of Financial Economics* 18, 373–399.

[5] Chen, N., Roll, R., & Ross, S. (1986). Economic forces and the stock market. *Journal of Business* 59, 383–403.

[6] Chih-Fong Tsai, Yu-Chieh Hsiao, "Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches", *Decision Support Systems* 50 (2010) 258–269.

[7] Cochrane, J. H. (1988). How big is the random walk in GNP? *Journal of Political Economy* 9, 893–920.

[8] Eshan et al., "Application of data mining techniques in stock markets – A survey", *Journal of Economics and International Finance* Vol.(2), pp.109-118, July 2010.

[9] Fama, E., & French, K. (1992). The cross-section of expected stock returns. *Journal of Finance* 47, 427–465.

[10] Fama, E., & French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33, 3–56.

[11] Ferson, W., & Harvey, C. (1991). The variation of economic risk premiums. *Journal of Political Economy* 99, 385–415.

[12] Galit Shmueli, Nitin R. Patel, Peter C. Bruce *Data Mining for Business Intelligence* 2010, Wiley, New Jersey.

[13] Hosmer, D. W., & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley.

[14] Javier Arroyo, Carlos Mat, "Forecasting histogram time series with k-nearest neighbours methods", *International Journal of Forecasting* 25 (2009) 192–207

[15] Jiawei Han, Micheline Kamber *Data Mining Concepts & Technique"* Second Edition (2000), Morgan Kaufmann Publishers, San Francisco.

[16] John E. Hanke, Dean W. Wichern *Business Forecasting,* Eighth Edition, 2005, Prentice Hall, New Delhi.

[17] K. Huarng, H.K. Yu, A type 2 fuzzy time series model for stock index forecasting, *Physica* A 353 (2005) 445–462.

[18] Kim, K.J.: Financial time series forecasting using support vector machines. *Neurocomputing* 55 (2003) 307-319

[19] Kumar, Manish and Thenmozhi, M. Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest, *Indian Institute of Capital Markets 9th Capital Markets Conference Paper.* [Online] Available: http://ssrn.com/abstract=876544 (February 06, 2006).

[20] Kyoung-jae Kim, "Artificial neural networks with evolutionary instance selection for financial forecasting" *Expert Systems with Applications* 30 (2006) 519–526

[21] Kyoung-jae Kim, "Artificial neural networks with evolutionary instance selection for financial forecasting", *Expert Systems with Applications* 30 (2006) 519–526

[22] Lakonishok, J., Shleifer, A., & Vishny, R. W. (1994). Contrarian investment, extrapolation, and risk. *Journal of Finance* 49, 1541–1578.

[23] Liao, S.-h., et al. Mining the co-movement in the Taiwan stock funds market. *Expert Systems with Applications* (2010).

[24] Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *Review of Financial Studies* 1, 41–66.

[25] Mark T. Leung et al, Forecasting stock indices: a comparison of classification and level estimation models, *International Journal of Forecasting* 16 (2000) 173–190.

[26] Menard, S. (1993). Applied logistic regression analysis, series: *Quantitative applications in the social sciences.* Thousand Oaks, CA: Sage.

[27] Muh-Cherng W, Sheng-Yu L, Chia-Hsin L (2006), "An effective application of decision tree to stock trading, *Expert Systems with Applications.*

[28] Myers, R. H. (1990). *Classical and modern regression with applications* (2nd ed.). Boston, Massachusetts: PWS-KENT Publishing Company.

[29] Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models* (2nd ed.). Homewood, IL: Richard D. Irwin, Inc.

[30] O'Connor, M., Remus, W., & Griggs, K. (1997). Going up-going down: How good are people at forecasting trends and changes in trends? *Journal of Forecasting* 16, 165–176.

[31] Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association,* 73, 699–705.

[32] Quah, T.S., Srinivasan, B.: Improving Returns on Stock Investment through Neural Network Selection. *Expert Systems with Applications* 17 (1999) 295-301.

[33] N.Ren M.Zargham and S. Rahimi "A Decision-Tree based Classification approach to rule extraction for security analysis" *International Journal of Information Technology and Decision Making* Vol. 5, No. 1(2006) 227- 240

[34] Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Ames, IA: The Iowa State University Press

[35] Studenmund, A. H. (1992). *Using econometrics: A practical guide.* New York: Harper Collins.

[36] Tuan Zea Tan, Chain Quek, and Geok See NG, "Biological Brain-Inspired Genetic complementary Learning for Stock Market and Bank Failure Prediction**,** *Computational Intelligence,* Volume 23, Number 2, 2007.