1.1 PREAMBLE

Modern Finance aims at identifying efficient ways to understand, transform, and visualize stock market data into useful information that will aid better in investment decisions. Financial tasks are highly complicated and multifaceted; they are often stochastic, dynamic, nonlinear, time-varying, flexible structured, and are affected by many economical and political factors (Tan et al 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 behavior of the market (Boginski et al 2006). 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. (Leung et al 2000). 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 of arbitrage. Therefore, being able to accurately forecast stock market index has profound implication and significance to researchers and practitioners alike (Leung et al 2000).
There are umpteen research work carried out suggesting the predictability of stock markets. Initially, tests of predictability of stock market returns were motivated by market efficiency, where it is assumed that predictability was inconsistent with constant stock market returns and efficient markets paradigm (David 2004). For long, it was thought that stock markets are not predictable, at least in an economically significant manner (Fama 1970). Lo & MacKinlay (1988) in their research paper claim that stock prices do not follow random walk and suggested considerable evidence toward predictability of stock prices. Basu (1977), Fama & French (1992), Lakonishok et al (1997) in their various studies have carried out many cross-sectional analyses across the globe and tried to establish the predictability of the stock prices. 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. While early studies suggested that such predictability was inconsistent with market efficiency and reflected inherent irrational investor behavior that may be exploited in trading strategies (Cutler et al 1989), predictability can also be viewed with respect to arbitrage pricing models (Ross 1976). 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 Research Motivation

Studying the behavior of stock markets is an interesting and challenging activity. The onus of studying the stock market behavior stems from the needs of forecasting, decision making, and risk management. The most popular Efficient Market Hypothesis (EMH) is often tested in order to
check the validity of the informational efficiency of stock market. The EMH asserts that no investor is in a position to make unexploited profit opportunities by forecasting future prices on the basis of past prices, yet a large number of researchers, investors, analysts and practitioners use different techniques to forecast or predict the stock index and prices.

Logistic regression (Hosmer & Lemeshow 1989, Press & Wilson 1978, Studenmund 1992 & Kumar & Thenmozhi 2006) and multiple regression (Menard 1993, Myers 1990, Neter et al 1985, Snedecor & Cochran 1980) are the conventional 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. Mathematical, stochastic, and econometric models have been used to model the share price behavior. But due to high volatility in the prices, the success of prediction (hit rate) is often in the low range. In order to reach better hit rates, researchers started employing data mining techniques such as neural networks 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 (Ren et al 2006). With the economy around the globe slowly recovering from the sub-prime crises, understanding the behavior of Indian stock markets during this post recession period will be helpful in understanding how far the Indian stock market is resilient in terms of its global counterparts. A data mining approach to Indian stock market analysis in comparison with other developed and developing markets will help us understand the predictability of the Indian stock market. Though there are many studies trying to predict the stock values and movements in the developed markets, very few studies are carried out in the Indian scenario to apply data mining tools. Even those few studied were trying to establish the supremacy of data mining tools over the statistical,
probabilistic and econometric time series models. Thus there arises a need for employing data mining tools to comprehensively study the predictability as a phenomenon. A study on predictability of Indian stock markets in comparison with other developed and emerging markets have not yet been undertaken. This gap motivates to quantify the predictability of both emerging markets and developed markets and to further analyzes how the predictability varies from long term to short term especially in the Indian context.

Another key gap is that there is no study in the Indian context to compare the efficacy of global cues as a predictor set over traditional technical indicators. Undeniable influence of global cues on the Indian stock markets made the researcher attempt to build predictive models with global cues as a predictor set.

As a piece of novelty, this study attempts to find out the co-movement of Indian markets with global cues by predicting the behavior of Indian Indices with frequently occurring patterns among the major times series of global indices by employing Apriori algorithm. This would help us to understand how global cues drive the Indian markets. Hence a research attempt is made to comprehensively analyze the predictability of Indian stock market indices with respect to other developed and emerging stock markets using a set of data mining tools.

1.3 BACKGROUND OF THE STUDY

High volatility in stock prices makes it difficult to predict the stock market movements. Hence, stock market analysts focus on developing models to successfully forecast index values or stock prices, aiming at high profits using well-defined trading strategies. The key to successful stock market forecasting is achieving best results with minimum required input data
The efficacy of the predictive models primarily depends on the input parameters chosen for the study.

The field of financial forecasting is characterized by data intensity, noise, non-stationary, unstructured nature, high degree of uncertainty, and hidden relationships (Hall 1994). The EMH claims markets are efficient and hence share prices move in random. However, academic research studies have proved that market prices are not random. Rather, they behave in a highly non-linear dynamic manner (Blank 1991). Modeling of stock index prices should consider the current mood of the market (bullish/bearish, boom/recession etc.) apart from other factors. There are myriad number of factors such as domestic and global economic conditions, political situation, investor’s psyche, catastrophe, and other unexpected events that influence the behavior of stock markets and it results in noise. The markets exhibit both linear- and non-linear behavior and hence models that can capture non-linear behavior are chosen for obtaining accurate predictability. Of late, various machine learning approaches are extensively employed to study the stock market behavior. Some of the popular machine learning approaches are decision trees, artificial neural networks (ANN), association rule mining, case based reasoning, support vector machines (SVM), clustering etc. Many studies have used ANN to forecast stock index movements. A large number of successful applications have shown that ANN can be a very useful tool for time-series modeling and forecasting (Zhang et al 1998). In particular, Back Propagation Neural Network is frequently used in stock market since the power of prediction is known to be better than others (Ming-Chi Lee 2009).

A neural network is more effective in describing the dynamics of non-stationary time series due to its unique non-parametric, non-assumable, noise-tolerant, and adaptive properties. However a critical issue concerning neural network is the over fitting problem. Neural network captures not only useful information contained in the given data but also unwanted noise
(Cao & Tay 2001a). Recent studies tend to hybridize certain ANN techniques (Heimstra 1995 & Witten & Frank 2011). The SVM is a novel neural network algorithm, developed by Vapnik and his colleagues (1998). The SVM implements the structural risk minimization principle. Also, the solution of SVM may be global optimum, and hence solves the problem of overfitting (Kim, 2003). Cao & Tay (2001b) examined the feasibility of applying SVM in financial forecasting and investigated the functional characteristics in forecasting. Lahmiri Salim (2011) applied the probabilistic neural network and SVM techniques to predict the stock market trend using technical and macroeconomic indicators and combining them together in order to compare the performance of the classifiers. Yang et al (2002), applied the SVM regression model to financial forecasting by incorporating market adaptions.

The most commonly used inputs are the stock index opening or closing price, as well as the daily highest and lowest values supporting the statement that soft computing methods use quite simple input data to provide predictions (Atsalakis & Valavanis 2009). Few studies have also used exchange rate of strong currencies along with daily prices of established markets like the Dow Jones, S&P (Ajith et al (2003), Huang et al (2007), Siekmann et al (2001), Wikowsta (1995) and technical indicators as predictors of stock index movements (Armano et al 2004). Wang et al (2011), proposed a hybrid approach combining ESM, Auto Regressive Integrated Moving Average(ARIMA), and Back Propagation Neural Network (BPNN) to forecast the stock index movement. Dai et al (2012) proposed a time series prediction model by combining nonlinear independent component analysis (NLICA) and neural network to forecast Asian stock markets. Ticknor (2013) has tried a Bayesian regularized ANN as a novel method to forecast financial market behavior using individual stocks. Most of the research is conducted using data from well-established stock markets as in the US, Western Europe, and Japan. In fact, only very few comprehensive studies are carried out to
understand the behavior of Indian stock market indices with data mining tools. Thenmozhi (2001) examined the feasibility of neural network in predicting the movement of the daily and weekly returns of BSE SENSEX Index. Pant and Rao (2003) in their work used ANN for estimating the daily return of the BSE SENSEX using randomized back propagation. Kumar and Thenmozhi (2006) have used Back Propagation Neural Network and compared it with a linear ARIMA model for forecasting different time series like INR/USD, Stock index return, index future returns etc. The results indicate that ANN-based forecasting method is superior to the linear ARIMA models. Kumar and Thenmozhi (2006) studied the predictability of S&P CNX NSE NIFTY index returns using SVM and Random Forest Regression. Hence, an attempt is made to investigate the predictability of the Indian stock market indices in comparison with developed and developing stock markets using data mining approach.

1.4 DATA MINING

Statistical tools have been applied for predicting the behavior of the stock market for more than half a century and a hit ratio of 54% was considered as a satisfying result for stock prediction. Analysts started employing more sophisticated mathematical, stochastic, and econometric tools to model the share price behavior. But due to high volatility in the prices, the success of prediction (hit rate) remained elusive. The advent of database, data warehouse technologies, and machine learning algorithms threw open hitherto unavailable possibilities and tools for the stock market research. These new data mining tools and techniques were found to outperform the classical statistical methods by 5–20% higher accuracy rate.

Data mining refers to extracting or mining knowledge from large amounts of data. According to the authors of the first book on data mining,
Frawlay et al (1991), data mining is “the non trivial extraction of implicit, previously unknown and potentially useful information from data…”

Another definition from the Gartner Group, the business research firm states that:

Data mining is the process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques. (Source: http://www.gartner.com/6.help/glossary)

Data mining involves techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, neural network etc. The major attraction of data mining lies in its ability to build predictive models apart from conventional retrospective models. In general, data mining tasks can be classified into descriptive mining and predictive mining. Descriptive mining tasks characterize the general properties of the data in the database whereas the predictive mining tasks perform inference on the current data in order to make prediction. Data mining tasks can also be classified into

- Forecasting
- Classification
- Clustering
- Associations
- Sequences
Forecasting and classification are essentially predictive techniques, while classification predicts categorical class labels like “bull” or “bear,” forecasting predicts continuous values. Clustering is a process of portioning or segmenting a given set of unlabeled data into many clusters so that homogeneity within the same cluster is maximized, while it is minimized between clusters. Frequent pattern mining searches for recurring relationships in a given data set, consequently may lead to the discovery of interesting associations and correlations between itemsets in transactional databases. Sequential pattern mining is the mining of frequently occurring ordered events or subsequences as patterns.

This study employs forecasting, classification, and association tasks of data mining to investigate the stock market predictability. Unlike statistical study, data mining does not involve formulation of hypothesis, sampling, and appropriate generalization that depends upon the finding of acceptance or rejection of the null hypothesis. Large databases and brute processing power of computers coupled with sophisticated algorithms enable data mining to process all the data involved in the study to come out with empirical findings.

Many of the popular data mining tools are machine learning algorithms that are able to learn the idiosyncrasies of the dataset as they process them. The more they process the dataset, the more they learn about the data. Typically, entire data under study is split into two groups, training dataset and test dataset. The training of the data mining models takes place with the training dataset. This learning phase is usually iterative and time consuming. The test dataset is earmarked to validate the performance of trained models. Once the model achieves the desired level of prediction accuracy with the training dataset, it is applied on the test dataset and the difference between predicted outcome and the actual values in the test dataset is measured. In data mining terminology, they are called prediction errors.
Using different error metrics available in the data mining literature, performance of the models, and predictability of the system under study can be evaluated.

1.5 RESEARCH OBJECTIVES

Exponential growth in processing powers of computers and their successful applications in solving complex mathematical problems has encouraged researchers to evince keen interest in the analysis and prediction of economical time series. Forecasting stock market returns and stock market behavior has received special attention. Though there are many traditional forecasting tools, with the growing size of stock markets around the globe and vast amount of data generated every day, sophisticated tools are employed to forecast stock prices or behavior. A plethora of research works have been undertaken addressing the predictability of stock market returns as well as the pricing of stock index financial instruments. Most of the proposed models rely on accurate forecasting of the level (i.e. value) of the underlying stock index or its return. In most cases, the degree of accuracy and the acceptability of certain forecasts are measured by the estimates of certain forecasts that are measured by the deviation of the estimate from the observed values. In the recent years, there have been a growing number of studies looking at the direction or trend of movements of various kinds of financial instruments (Wu & Zhang 1997; O’Connor 1997). However, these studies did not provide a comparative evaluation of different classification techniques regarding their ability to predict the movement of the stock index return (bullish/bearish). Hence, the following research objectives are formulated to empirically investigate the predictability of stock index movements and compare the efficacy of various models.
1. To predict the movement and price of select global stock indices using k-Nearest Neighbor (k-NN), ANN, decision tree, and SVM algorithms.

2. To compare the performance of above models in predicting the stock index movement and price.

3. To study the efficacy of various global cues in forecasting the Indian stock index movement/price.

4. To study the relationship between Indian stock market and global cues.

5. To generate predictive rules for the movement of Indian stock indices by association rule mining using Apriori algorithm.

1.6 RESEARCH TOOLS

The current research aims at investigating the predictability of the stock market indices using various data mining tools. Popular machine learning algorithms such as ANN, k-NN, SVM, decision tree and its ensemble variant called Random Forest are employed in this study to investigate predictability of global and Indian stock market indices. Brief descriptions about the tools are given below.

ANN: It is an efficient information processing system that resembles the characteristics of a biological neural network. The ANNs possess large number of highly interconnected processing elements called nodes. Each connected link is associated with weights that contain information about the input signal. This information is used by the neural net to solve a particular problem. The collective behavior of the ANNs is characterized by their ability to learn, recall, and generalize training patterns or data similar to that of a human brain.
**k-NN algorithm:** It is a data mining prediction technique that looks for records with similar predictor values in the historical database and uses their prediction values for making prediction about an unclassified record. As the name implies, this algorithm looks for “k” nearest cases in sample dataset and uses outputs of those nearest cases for forecasting the output for a new instance of data. Here, “k” is usually a small integer that can be optimized by various trials and errors during the training phase of the model. The output of all neighbors can be given equal weights or the closest neighbors can be given more weights with inverse proportion to its distance from the new data record. One of the most widely used metrics for identifying nearest neighbor is Euclidean distance.

**SVM:** It is a promising new machine learning algorithm developed by Vapnik in 1995 that can be used to forecast both linear and nonlinear data. It uses nonlinear mapping to transform training data into higher dimension plane. Within the new dimension, it searches for a linear optimal separating plane, demarcating one class from the other. The transformation of data from lower dimensional to higher dimensional plane is done by kernel functions. The SVM tries to create an optimal line of separation or hyperplane between the objects of two different classes. It also tries to optimize the width of the hyperplane with two boundaries that is known as maximal margin hyperplane. The instances that touch the margins are support vectors.

**Decision tree and random forests:** Decision tree is based on constructing hierarchical trees from the historical dataset available and fully grown and pruned trees are then used for forecasting new instances. Decision trees are represented by a set of questions that splits the entire dataset into smaller and smaller subgroups. It searches for one predictor variable and its particular value that splits the entire dataset into two parts with maximum homogeneity in terms of decision variable. The choice of split variable is based on impurity
function. The predictor with the least impurity measure is the choice of splitting the dataset. This splitting process is then continued with each of the resulting data fragments until “leaves” or decision nodes are reached. The resulting tree can be used for making decision rules for forecasting.

Random forests are an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. This study employs random forest as the classifier.

Apriori algorithm: It is a classic algorithm proposed by R. Agarwal and R. Srikant in 1994 for mining frequent item sets in transactional databases. The name of the algorithm is based on the fact that it uses prior knowledge of the frequent item set properties. It proceeds by identifying the frequent individual items in the database that satisfy minimum support level. It then employs an iterative approach by extending them to larger and larger item sets until no more frequent item sets are found in the database. The frequent item sets determined by Apriori lead to the discovery of association rules and correlation among the items in large transactional or relational data sets.

1.7 DATA AND SAMPLE SELECTION

Data required for the study is obtained from the respective websites of various stock exchanges. The daily stock prices of various indices for the period from January 1, 2006 to May 31, 2013 are collected. The major stock indices of India (BSE SENSEX, NSE NIFTY), stock indices of developed countries including America (NASDAQ & NYSE), Canada (GSPTSE), England (FTSE), France (CAC40), Germany (DAX), Japan (NIKKEI) and
stock indices of developing BRIC countries (IBOVESPA, RTSI, SSE, HSI) are chosen for the study. The stock markets are chosen on the basis of the total market capitalization and the most widely followed indices in the respective markets are chosen for the study.

Eleven major global cues comprising of daily close price of NASDAQ, NYSE, NIKKEI, FTSE, daily exchange rates of the strong currencies such as US Dollar, Pound, Euro and Yen, daily gold price, Brent oil price, and LIBOR rate are considered as the global cues. They are chosen on the basis of the degree of association they exhibit with the Indian stock indices.

1.8 METHODOLOGY

The entire data set containing daily open price, high price, low price, and close price of the index for the period of seven and a half years (Jan 1, 2006 to May 31, 2013), comprising of more than 1800 trading days. Along with the intra-day movements of index price, select technical indicators are computed and they form the input variables for the predictive models.

Data reduction aims at achieving parsimony and eliminating multicollinearity problem. Principal component analysis reduces the number of input variables to the predictive models by grouping the highly correlated inputs into a single set. The inputs comprise of intra-day movements of the index such as opening price, high price, low price, closing price, and widely used technical indicators such as stochastic %K, rate of change of price (ROC), moving average convergence and divergence (MACD), price minus moving average (P-MA), accumulation and distribution oscillator, price oscillator (OSCP), and relative strength index (RSI).
The entire data set is split into training dataset and test dataset in the ratio of 80:20, respectively. Data for the period from January 2006 to December 2011 (~1500 trading days) constitute the training data set, and the remaining data from January 2012 to May 31, 2013 (~350 trading days) constitute the test dataset. Predictive models are built and trained using the training dataset employing four data mining algorithms. These models are then put to actual use to predict the closing value of the next day’s index price of the test dataset. The actual closing values in the test data set and the predicted outputs from the models are compared and the prediction accuracy is evaluated using the performance metrics. The performances of predictive models are evaluated using the following statistical metrics: correlation coefficient, mean absolute error, root mean squared error, relative absolute error, and root relative squared error. The same procedure is adopted for predicting the next day’s movement using classifier models that predict the class of the outcome – either bull or bear. Hit ratio, error rates, and Kappa statistics are the performance indices used to evaluate the performance of classifier models.

The famed Apriori algorithm of market basket analysis is employed to study the relationship among Indian indices and global cues. The daily price of NSE NIFTY, NASDAQ, NYSE, NIKKEI, FTSE, exchange rates of the strong currencies such as US Dollar, Pound, Euro and Yen, gold price, Brent oil price, and LIBOR rate are captured as date-wise transactional record. The closing price of all the major indices and values of global cues are coded as “1” if today’s closing price is higher than yesterday’s closing price, otherwise “0”. Apriori algorithm employs iterative process to identify frequent item sets with a minimum support of 10%. From those frequent item sets, strong association rules are generated satisfying a confidence level of 80%.
Weka’s implementation of above mentioned algorithms are used in this study. Weka is one of the widely used open source software, popular among the data mining community. It was developed at the University of Waikato in New Zealand and the name is an acronym for Waikato Environment for Knowledge Analysis.

1.9 CONCLUSION

This empirical study is an attempt to measure the predictability of Indian and Global indices using the data mining approach. By employing four popular data mining tools predictability of the indices are measured in terms of value and direction. Predictability of emerging and developed markets is compared. The study reveals that emerging markets are less predictable than the developed markets. The results of the study prove the emergence of SVM as the most accurate forecasting and classifier model. Term-wise analysis of the predictability on Indian indices shows that markets are better predictable in short term than in the long term. The impressive growth in hit ratios of the classifiers and drastic reduction in forecasting errors, as one move from long term to short term prediction, is observed in the results. The study also highlights the effect of various global cues on Indian indices. These global cues, when used as a predictor set for predictive models, in lieu of the set of technical indicators, is found to result in a better predictive performance of the models. Association rule mining help identify patterns among the global cues that would result in bullish and bearish trends in the Indian indices. These association rules would be of immense help in predicting the trend of Indian markets. Thus this study throws more light on the dynamics and behavior of Indian stock markets.
1.10 ORGANIZATION OF THE THESIS

As this introductory chapter delves about the background, need, and objectives of the study, Chapter-2 discusses extensive literature review and various data mining approaches to stock market prediction. This also explains the methodology adopted for the research, explaining the data, sources of data, and tools employed in the study.

Chapter-3 is devoted exclusively to forecasting global and Indian indices, trying to compare the predictability of developed and emerging markets. It then attempts to find out the best forecasting tool. It tries to explore the variation in the predictability of Indian markets in different time frames. It also tries to find whether global cues will help forecast better than the widely used technical indicators.

Chapter-4 is the application of classifier models to predict the movement of global and Indian indices. It then attempts to find out the best classifier tool from the four tools employed in this study. Like the forecasting models, it tries to explore the variation in the predictability of Indian markets, in terms of hit ratio, in different time frames. It also tries to find whether global cues form a better predictor set for classifier models than popular technical indicators.

Chapter-5 deals with the relationship between global cues and Indian markets and applies the famed Apriori algorithm for association rule mining.

Chapter-6 highlights the summary of findings, conclusions, and future scope of this research study.