A Data-Integrated Tree-Based Simulation to Predict Financial Market Movement

Durai Sundaramoorthi, Washington University in St. Louis, USA
Andrew Coult, Missouri Western State University, USA
Dung Hai Nguyen, Washington University in St. Louis, USA

ABSTRACT
The Standard and Poor’s 500 Index (S&P500) is one of the commonly used indices on the New York Stock Exchange. The 500 publicly traded companies that make up the index are chosen by a committee to best reflect the overall market of the United States. The broader objective of this research is to estimate the dynamics of the financial market movement in the United States. It is achieved by developing a data-integrated tree-based simulation model to predict S&P500 open and close values for a week. Classification and Regression Trees (CART) - a data mining method - is utilized to extract patterns of the financial market dynamics based on a data set collected from May 1, 2008 to November 30, 2009. The data set included the daily movement of financial markets in seven countries in Asia and Europe in relation to the daily movement of the S&P500. CART also utilized data on the currency exchange rates to capture the financial dynamics between the US and other countries. The simulation model repeatedly samples from four trees developed by CART to know how the opening and closing values of the S&P500 move in tandem with the other markets.

Keywords: Classification and Regression Trees, Data-Integrated Tree Based Simulation, Financial Market, Financial Market Movement, Simulation

INTRODUCTION
The Standard and Poor’s 500 Index (S&P500) is one of the commonly used indices on the New York Stock Exchange. It was introduced in its current form in March, 1957. The 500 publicly traded companies that make up the index are chosen by a committee to best reflect the overall market in the United States. The S&P500 hit its highest point on October 9, 2007. Since then, the United States economy has been on a generally downward trend, and so by extension, has the S&P500. This collapse began with the financial sector, stemming primarily from subprime mortgage problems, but spread rapidly into other sectors of the economy. The S&P500 was characterized by unusually high volatility in 2009. The recent apparent unpredictability of the market has resulted in the loss of millions of dollars of investments.
One of the primary tasks of institutional fund managers and financial analysts is to predict how the market is going to move on a daily basis so that they can better reach their returns goals. To aid the fund managers and financial analysts, this research develops a data-integrated tree-based simulation model to predict how the opening and closing price of the S&P500 move in tandem with the other markets. Classification and Regression Tree (CART) model – a data mining tool for prediction and classification (Breiman, Friedman, Oishen, & Stone, 1984) - is used to develop four regression tree structures: (1) “first” tree predicts S&P500 Open value for Monday mornings based on other market indices, currency exchange rates, and S&P500 open and close values of the prior Friday; (2) “second” tree predicts the S&P500 close value for Monday evenings based on other market indices, currency exchange rates, S&P500 open and close values of the prior Friday, and the predicted S&P500 open value for that morning; (3) “third” tree predicts the S&P500 open values for Tuesday through Friday based on previous day’s predicted S&P500 open and close values; and (4) “fourth” tree predicts the S&P500 close values for Tuesday through Friday based on the predicted value of that day’s S&P500 open value, and previous day’s open and close values. Simulation models developed by sampling from these four trees are better representation of the actual system and more efficient to execute.

**Contribution**

There are two major contributions made in this research:

- This research introduces a novel approach to the finance modeling community for constructing efficient simulation models based on data mining. This way of simulation modeling avoids misrepresentation of financial dynamics and characteristics because it is entirely based on the pattern learned from a real data set collected from the financial market over a long period of time. Moreover, this approach reduces simulation states and is consequently more efficient to run.
- This research introduces a tool to predict S&P500 open and close values for one week in advance. The simulation model enables fund managers and financial analysts to make investment decisions based on a scientific evidence-based approach.

The rest of this paper is organized as follows: First, we provide a literature review on financial market modeling, data mining, and simulation. Afterwards, a brief introduction is given on data. Then, we describe the data mining tree structures used to build the simulation model. Following that, we present S&P500 predictions and their comparison with the actual data. Finally, we provide concluding remarks including discussions on a possible simulation-based optimization approach to optimize a portfolio.

**LITERATURE REVIEW**

There are three major components in this research: finance market modeling, data mining, and simulation. This section gives a brief literature review on each of these topics.

**Finance Market Modeling**

The early financial market models assumed market is characterized by a random walk (Griffin, 1977; Beaver, 1970; Box & Pierce, 1970; Wu, Kao, & Cheng, 1996). Autoregressive integrated moving average (ARIMA) model was one of the earliest techniques used for time series data in econometrics. Especially, Box-Jenkins method (named after statisticians George Box and Gwilym Jenkins) and its variants are popular to this date for applications with stationary or nonstationary time series. Traditionally, differencing approaches or transformations are used to achieve stationarity in non-stationarity time-series (Box & Jenkins, 1970). Over the years, several studies further refined ARIMA models and invented new models to handle different situations. Some of those models are dynamic
Related Content

An Excel-Based Decision Support System for Supply Chain Design and Management of Biofuels
[www.igi-global.com/article/an-excel-based-decision-support-system-for-supply-chain-design-and-management-of-biofuels/120446?camid=4v1a](www.igi-global.com/article/an-excel-based-decision-support-system-for-supply-chain-design-and-management-of-biofuels/120446?camid=4v1a)

Managing Fisheries in Light of Complexity and Chaos Theories
[www.igi-global.com/chapter/managing-fisheries-light-complexity-chaos/70893?camid=4v1a](www.igi-global.com/chapter/managing-fisheries-light-complexity-chaos/70893?camid=4v1a)

Capacity Management Business Processes
[www.igi-global.com/chapter/capacity-management-business-processes/6095?camid=4v1a](www.igi-global.com/chapter/capacity-management-business-processes/6095?camid=4v1a)
Minimizing Makespan on Identical Parallel Machines
www.igi-global.com/article/minimizing-makespan-on-identical-parallel-machines/124759?camid=4v1a