Journal of Islamic Monetary Economics and Finance, Vol. 5, No.3 (2019), pp. 475-490
p-ISSN: 2460-6146, e-ISSN: 2460-6618

BIG DATA ALGORITHMS AND PREDICTION: BINGOS AND RISKY ZONES IN SHARIA STOCK MARKET INDEX

Shahid Anjum[1] and Naveeda Qaseem[2]

[1] School of Business Universiti Teknologi Brunei, Brunei Darussalam, anjumsw@hotmail.com
[2] University of Westminster, United Kingdom, N.Qaseem@westminster.ac.uk

ABSTRACT

Each country with a stock exchange normally calculates various indexes. So is the case for Malaysia's Kuala Lumpur Stock exchange (KLSE). FTSE BURSA Malaysia EMAS Sharia price index (FTBMEMA) is one of its Sharia indexes. In an effort to find which other indices may forecast this Sharia index, we selected 23 relevant indexes and two exchange rates. Momentum indicators for short, medium and long term have been calculated for the variables. The objective of this study is to find predictive indicators for FTBMEMA out of the population of 188 original and derived variables. Difficulty arises in reducing the number of variables for regression or other predictive models like neural networks. In this preliminary study, data mining attribute selection algorithms along with cross validation criteria have been used, through the use of Java class library Weka (JCLW), for reducing the number to statistically relevant variables for our regression estimation in an effort to forecast various performance parameters for FTBMEMA like performing either in a mean performance range, having jackpots and bingos or falling into danger zones. Provided the extent of the required predictive accuracy, the results may bring additional insights for diversifying and hedging various types of investment portfolios as well as for maximizing returns by portfolio managers.

Keywords: Sharia Stock Market Index, WEKA Class Library, Big Data Mining, Attributes Selection, Prediction Analysis.
JEL Classification: C45; D53; E37; G17.

Article history:
Received : December 07, 2018
Revised : May 3, 2019
Accepted : September 29, 2019
Available online : November 1, 2019

https://doi.org/10.21098/jimf.v5i3.1151

I. INTRODUCTION

Starting from the last years of the eighties and accelerating throughout the nineties of last century, international landscape has changed and is continuously changing substantially and multi-dimensionally. The underlying factors behind this international development are many, ranging from historic events to technological innovations. Financial markets are the most efficient (in the economic sense) and most dynamic entities; therefore, their integration across the globe was also fast paced. Though trade and investment functions led the dynamics of globalization, financial function being a complimentary function was also a vital part of the globalization efforts. This is because as the world financial markets have started their journey towards globalized economic markets and possible financial integration through financial liberalization, currency as well as banking crisis has become more frequent. The proliferation of financial crisis whether banking crisis or currency crisis or both, has become a frequent occurrence in present days integrating financial and trading world.

Stock market investments are more risky as compared to other form of financial investments because the share price movements are not easy to forecast and stock trading, therefore, is highly a speculation activity. There are various ways to predict the share price movement in order to generate alpha by beating the market, but predicting the stock market is like a mission impossible as there are remote possibilities of accurately predicting the movement of a specific company's share prices over shorter period. The prediction of the movements of share or index prices has been a domain of various disciplines and has been extensively studied by finance, economics, statistics, mathematics, insurers, and computer professionals who have applied various forecasting techniques, formulas, formulations, and algorithms on the share prices' historical data and using the concepts from fundamental and technical analysis (Edward & Magee; 2001 and Jones; 2007). New applications of information technologies to financial domain, i.e., FinTech, are exploiting the prediction dimension of stock price through the approaches like artificial intelligence (AI), data mining and recommender systems. In this context various stock markets' style indices and indices of other allied financial and commodity markets like oil and precious metal futures, forex, and allied regional indices are assumed to be correlated but still may be holding valuable information regarding the price movements of the other indices. The focus of this study is to use the selected indicators or attributes selected through attribute selection algorithms (see Anjum et al., 2017c), which are important to predict index and are tested using simple regressions.

Twenty-five variables were cleansed and chosen to be transformed into further derived variables based on the insights provided in financial technical analysis literature. We want to sort out the relevant indicators for our target variable using data mining algorithms and then based on results of the selection, the aim is to recommend the important ones which can be used for further prediction either using various data mining classifiers, rule based or clustering algorithms, neural networks, Bayesian networks, deep learning, or simple regression analysis techniques (Tjung et al.; 2010). This exercise left the dataset comprising of one hundred and eighty eight independent variables. As we know that most of the technical indicators based transformations of twenty-five variables may be

redundant and may not have any effect on our dependent variable, which is 10 days future FTSE Bursa Malaysia EMAS Sharia price index. The process of data mining typically consists of the following three steps: data preprocessing, such as cleansing and others, data analysis through the application of various algorithms, and representation and interpretation of results.

By using the data mining algorithms with the help of JCLW and with the help of attribute selection approach, we have narrowed down the variable space for the prediction of the future value of the target Sharia stock index. Moreover, by making use of appropriate weights, normalization schemes, the study made use of ranking algorithms to sort out a subset of statistically relevant variables from amongst all of the 188 variables in the dataset. Using data mining algorithms, the study successfully reduced the number of variables to eighteen, used for regression estimation - our predictive methodology. The selected, reduced number of variables may also be used for various other predictive methodologies like neural networks and deep learning methodologies, which this study may be improved on in future. Our study pioneered in a completely different approach for variable reduction, considered for selection of statistical significant variables for regression analysis. Previous studies used hit and trial method of variables selection. The proposed new method in this study has proven to be a time saver as well as based on solid scientific footing.

The paper is divided into five sections. After this section of introduction, we will describe a brief review of literature about the issues in the financial and stock markets in sub-section 2.1, which is followed by sub-sections 2.2 and 2.3 for brief review of literature on data mining as well as big data and financial indices respectively. Section 3 discusses data description and methodology and section four provide details on indicators' selection, validation and regression results in sub-section 4.1 as well as analysis in sub-section 4.2. Finally, section five provides the conclusion of the study and recommendations for future research.

II. LITERATURE REVIEW

2.1. Financial and Stock Markets

Financial sector, in general, consists of sub-sectors like banking sector, public finance sector, insurance and other institutional saving sector and corporation sector. Such sectors consist of different type of markets like debt market, equity and stock market, bonds market, commodities, and derivatives market and foreign exchange (forex) or currencies market. Various categories of stock market style indices are calculated for the financial markets players' use, including, among others, small-cap, mid-cap and large-cap growth indexes and, small-cap, mid-cap, large-cap value indexes, socially responsible indexes for investments (SRI) categorized by country and global ethical stock market indices. As each country with a stock exchange usually calculates not only the index representing the entire stock market, but also various sectoral and stylized indices as well, it is, therefore, important to find the inter-relationships among various types of the index categories, representing whether various stylized indexes of various sectors within a country, across various financial, derivatives and commodity markets, or various indexes in the regional and global markets. This is a highly value-added

exercise because based on efficient market hypothesis, the indices are assumed to carry all types of information required for investment decision making which is reflected in the index values, which change dynamically throughout during the trading activity in the market.

This complexity of financial markets and in modeling them has been reflected in various recent financial crisis. Although the worst financial crisis was triggered by mortgage sector in USA in 2007-08, but Asian financial crisis of 1997-98 has also shown the interdependent nature of financial as well as real markets in ASEAN region. A study by Cooper et al. has summarized the views of the Asian Crisis into four categories. First, one is the weaknesses of industrial, financial and exchange rate policies within Asian economies. Second is the over investment based not on economic rationale rather than based on moral hazards originating from IMF bailouts' money and cronyism. Third and fourth are exchange rate devaluation and creditors' panic (financial) in the middle of Asian financial crisis of 1997 in Thailand. Because of liberalization of financial sector and capital account in the late eighties in Asian economies, huge inflow of private credit by foreign lenders has taken place. The first and second views are based on this background. Cooper et al. (1998) has, however, disagreed with the two views. The third view came from the background that foreign reserves had depleted and that central bank could not maintain the promise of maintaining the fixed exchange rate regime. Investors, therefore, had forecasted this devaluation and exited the country before this regime shift. The fourth view based on creditor's panic, in the stock markets mainly, has been advocated as the core reason for the crisis. This view actually has argued that large short-term foreign currency dominated debt has sky rocketed resulting in foreign reserves depletion in Asia. In short, there was a shift in the market expectations, which have been reinforced by the belief that central bank's ability to defend against the attack was limited. This has led to self-fulfilling and contagious attack. Despite the economic fundamentals in Asian economies being strong enough to ensure the servicing of the long-term foreign currency denominated debt without the possibility of default, the crisis otherwise occurred. Lack of adequate banking supervision as well as collusive relationships between big businesses and the government equally aggravated the panic by specific Asian institutional conditions. The creditors overrreacted to both the positive as well as negative news. The issue of contagious nature of the crisis proved to be more dangerous because of two issues. First, it has to do with how one market's volatility transmitted from one market to another market within the same economy, i.e., cross-market level volatility, has not been addressed. Second, cross-country or broader level contagion effects have not been addressed and nor the channel(s) through which it occurs. Such a whole story tells the tale of not only how complex financial markets are, but also how interconnected they are and on the top of their impact are the shocks coming from the stock markets. Stock markets are not only volatile, but also hard to model and predict, therefore, panic created by even rumors may create a lot of damage (Cooper and Bosworth 1998).

Stock markets are unpredictable by nature. In order to estimate and predict the movement of share prices by making use of the knowledge domains of statistics, mathematics and forecasting methods, a study using historical share price data (named as technical analysis (TA)) has been around for 100 years. Rockefeller

(2004) reported the study and the professionals who had access to real time stock prices data conducted it. Uncertainty is the main source of risk in financial business and, therefore, the focus of profit savvy investors who are focused on maximization of their returns. Uncertainty is also one of the factors involved in public risk perception as described by Fiksel and Covello (1987). Meanwhile, Tirea et al. (2013) described Wave Analysis stock prediction model, which extended stock analysis by adding agent, based system and has used Fuzzy Logic Theory, Neural Network methodology, and Elliott Wave Theory. It was developed by Atsalakis et al. (2011), who developed a framework (Stock Market Multi-Agent Recommendation System - SMMARS). The study combined the results of the Technical Analysis Methods with the Neural Network Methods (Multi-Layer Perceptron) into a Multi-Agent Architecture for better forecasting the future trend of stocks at the exchange. SMMARS used some special technical analysis methods, in order to point out good buy/good sell moments and to better forecast market trends in the Bucharest Stock Exchange Market (BSE).

Avery et al. (2015) studied approximately 5.0 million stock picks submitted by more than 60,000 individual users from November 1, 2006 to December 31, 2011 to the "CAPS" website run by the Motley Fool company. The CAPS uses collaborative filtering proven useful in a wide variety of contexts. A Fama-French decomposition suggests that stock picking rather than style factors largely produced the results. The result of this analysis suggested that stocks with high ranks by Motley Fool participants earned higher subsequent raw returns and vice versa. If that were true, the historical success of momentum strategies could explain the return patterns. Historical data is required for the understanding about the investment products and to recommend the right products in which the client might be interested to invest in.

2.2. Data Mining

The investment domain is a dynamically changing one, stochastic in nature and highly unpredictable environment. All of this demands of portfolio manager to create the investment portfolio based on efficient analytical tools. The investment banks and financial institutions increased their reliability on the availability of big data for stock markets. Finding correlations in historical data of a stock in order to analyze current and future trends before making a buy or sell decision is a commonplace norm. Stocks price movements being 'random walk' processes and the fact that prediction might be run off due to some unexpected news with impacts on the respective company have made the stock market predictions a daunting task (Reilly & Brown; 2012). The obvious complexity of the problem paved the way for intelligent prediction paradigms (Yong et al. 2009) in the stock markets. The art of modeling is a highly scientific process. There are seven principles of strategic thinking as defined by new science that is helpful for modeling any phenomenon (Sanders; 1998). There are various ways to predict the movement of share prices in order to generate alpha by beating the market but predicting the stock market is like a mission impossible as no one can accurately predict the movement of a particular share prices for company over shorter period. Prediction of the movement of the share prices was a domain of various disciplines and has

been extensively studied by finance, economics, statistics, mathematics, insurers and computer professionals who have applied various forecasting techniques, formulas, formulations and algorithms on the share prices' historical data and using the concepts from fundamental and technical analysis.

New applications of information technologies to financial domain, i.e., FinTech, exploit the stock price prediction dimension through the approaches like artificial intelligence (AI), data mining and recommender systems. Using data mining techniques like attribute selection, clustering, classification, neural networks and association rules has helped quantitative analysts to determine correlations between the trend a stock follows and the reaction of a customer to the change in trends which may provide some indications of the behavior of stock indices as well. The process of discovering hidden patterns, trends and extraction of knowledge from large data sets is known as data mining. New rules are extracted through uncovering unknown patterns from large databases between objects are potentially useful in making crucial business decisions. Using data mining techniques like classification, attribute selection, clustering, association rules and various kinds of network schemes (refer to Table 1) have helped quantitative analysts to determine correlations between the trend a stock follows and the reaction of a customer to the change in trends. Data mining is an important knowledge discovery system being used in different data intensive systems. Recently data mining approach has already started a benchmark regarding the efficiency of this technique in the financial applications over the conventional techniques.

Table 1. Different Data Mining Techniques and their description

Concept Description	Characterization and Discrimination
Attribute Selection	Java Class Library Weka (JCLW) name: weka.attributeSelection.... Evaluates the relation of a subset of factors by considering the individual predictive ability of various features
Association	Correlation and causation based relationships
Classification	Construct models (functions) to distinguish concepts for prediction and classification schemes can be decision-tree, classification rules, or neural networks
Numeric Prediction	Prediction of unknown or missing numerical values
Cluster analysis	Group data to form new classes by finding distribution patterns, maximizing intra-class similarities and minimizing inter-class similarities
Outlier analysis	Analysis of a data object which falls in the category of not-normal behavior, noise or exception?
Trend & evolution analysis	Detecting trends and deviations with the help of regression, periodicity or similarity-based analysis and sequential pattern mining
Other Algorithms	Pattern directed analysis or statistical analyses techniques

Source: By the Author from different sources

Data mining approach not only has a variety of methodologies available to fit different problems, but also can handle different types of data sets. It can also handle the nominal, categorical, Boolean and binary data besides common numerical form of data. In data mining methodologies, several data types are used. If the data type

is of attribute type, then attribute-value approach is used which in turn can be either statistical or connectionist (Breiman; 2001). In addition, data sets can be time series or a kind of all other variables types that may influence the evolution of the time series. Data Mining assumes the functional form of relationships being modeled. Specific aspects of the data mining technique are knowledge investigation and creation techniques, processes, algorithms and mechanisms from data stocks as stated by Norton (1999). Langdell (2002) described the examples of the data mining techniques being used in the financial applications. Data mining process can be divided into five stages, i.e., understanding the problem, defining the objective, preparing the data (data acquisition and entry, data integration and making of final data sets), mining the data, and presenting the results. In order to achieve its objective, the study made use of the data mining algorithms with the help of Java class library Weka's attribute selection approaches provided in studies like (Bouckaert et al. (2016), Kovalerchuk and Vityaev (2000), Langdell (2002); Yadav et al. (2014) and Hart et al. (2010).

2.3. Financial Indices and Big Data

The type of available data to describe items may take the form of an unstructured data, a semi-structured data or a structured one. Such a data feature has multi-faceted impacts, e.g., on algorithms and on user models. Real-life data typically needs to be preprocessed, e.g., cleansed, filtered, transformed, in order to be used by the machine learning techniques in the analysis step and this preprocessing of our dataset has been taken on in the section below. Amongst all the processes in data mining, data cleansing task is not only cumbersome but also time consuming which are estimated to be around over 35% of all work in data mining as per various studies in empirical data mining literature. This is also the focal point for our current research as we want to have the right set of Sharia stock index predictor variables at hand but also want to narrow down this space for practical implementation as well. We want to sort out the relevant indicators for our target variable using data mining algorithms and then based on results of the selection, the aim is to recommend the important ones which can be used for further prediction either using various data mining classifiers, rule based or clustering algorithms, neural or Bayesian networks or simple linear regression. Predictive variables for the market returns especially in the macroeconomic category along with their references survey has been presented in a tabulated format in Anjum et al. (2017b). The list of variables includes economic fundamental, macroeconomic variable, inflation, money supply, ASX banking and finance index, consumer price index, seasonally adjusted, foreign portfolio investment, consumption-wealth ratio, exchange rates, consumption for US, price earnings ratio, industrial production, exports and oil prices. The weak form of efficient market theory postulates we have, therefore, collected the data for Malaysian stock markets' various style indices, foreign exchange market and other allied financial markets regional indices like commodity and precious metal futures markets only for our technical analysis (TA) based prediction for Malaysian Sharia Index. In classical finance, Fama and French (1993) identified three measures that have been demonstrated to predict future stock returns. We use the three factors identified by Fama and French in

our analysis along with the fourth technical factor, momentum (Mom) which has been identified by Carhart (1997) (Avery et al. 2015). The same TA method, however, does not work in the same stock analysis all the time (Rockefeller, 2004). We assume this even further for data from various different countries or even for different periods.

III. METHODOLOGY

3.1. Data

We have started with the data for twenty-five independent variables and a dependent or target variable (or class in data mining terminology), as have been discussed in section 2.3 above. After cleansing, all the variables have 2550 observations of daily prices and/or indices, which comprise of the period from October 2006 until August 2016. Amongst the independent indicators included are twenty-five variables which are spot values as well as lagged variables which are five, ten, twenty, thirty and fifty days lags of various Malaysian stock market indices. Lagged variables have been taken for variables like FTSE BURSA Malaysia price indices e.g. KLCI (FBMKLCI), Top 100 (FTBM100), Mid 70 (FTBMM70), Small Cap price index (FTBMSCP), Fledgling (FTBMFLD), FTSE/ASEAN 40 index (FTAS40I), Hijrah Sharia (FTBMHJS), EMAS Sharia price index (FTBMEMA) and ACE (FTBMMES), AI POI PLTN (FTBMAPL), Consumer GDS (FTMCGL-PI) and Malaysia-DS Con & Mat (CNSTMMY) indices, Kuala Lumpur price indices e.g. SE Trade & Service (KLSETAS), SE Industrial Prod. (KLSEINP), SE Industrial (KLSEIND), SE Technology (KLSETEC), SE Construction (KLSECON), SE Consumer Prod. (KLSECOP), SE Finance (KLSEFIN), SE Properties (KLSEPRP), SE Plantations (KLSEPLN) and SE Tin & Mining (KLSETIN) as well as closing prices of crude oil future (Singapore), commodity research board index (i.e., Cmd R Brd Index or CRBI), FTSE Straits Times index (GMC-STI-SNG), gold spot price index and exchange rates (MYR versus US$ & MYR versus SNG$). We have used FTSE BURSA Malaysia EMAS Sharia price index (FTBMEMA) as the target variable (or class in data mining terminology) and known as independent variable in regression domain while all other variables as indicators. The definition of various nominal categories in the stock market returns for FTSE BURSA Malaysia EMAS Sharia price index (FTBMEMA hereafter), which have been used in this study, have been based on the concepts presented in the studies like Anjum (2003), Radelet & Sachs (1998) and Anjum (2017c).

After having the exhaustive dataset, we have focused on narrowing down the variable space to make the dataset of reasonable size for the prediction of the future value of the target variable Sharia stock index. This has turned us towards the focus of our research, which is to select and recommend the most relevant explanatory variables for the prediction of this target variable by eliminating the redundant or ineffective variables by using data mining algorithms.

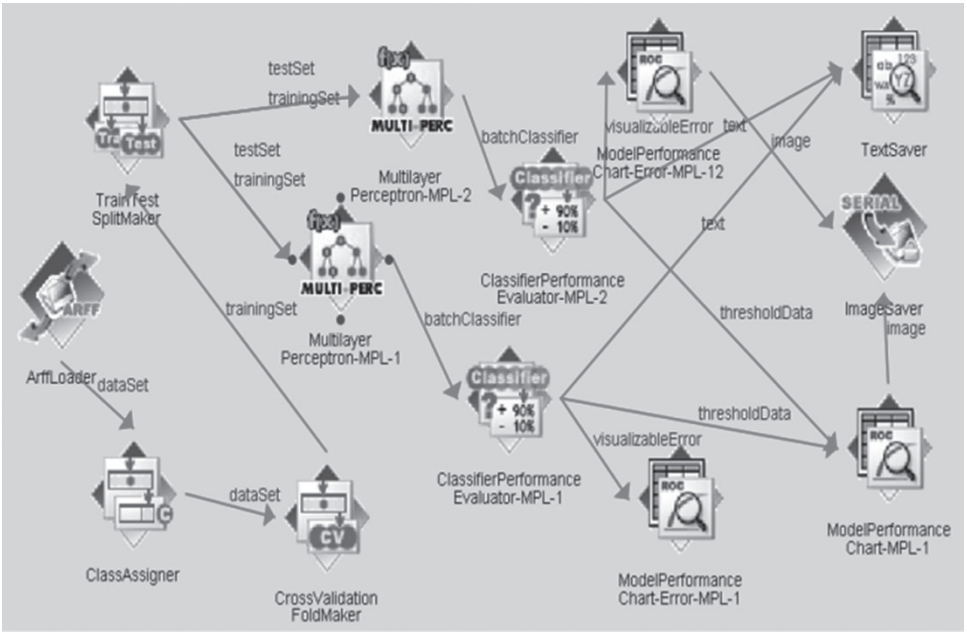
3.2. Method

There are various kinds of models available in the context of measurement of relationships among various variables. The models may belong to any category of

a combination of four dimensions i.e., stochastic or deterministic dimension versus conceptual or empirical dimension thus making four distinct classes of models, which are named, as stochastic-conceptual, stochastic-empirical, deterministic-conceptual and deterministic-empirical. Any of the main classes may be sub-classified in several ways like linear or non-linear in the systems-theory sense, linear or non-linear in the statistical regression sense and lumped or probability-distributed or geometrically distributed. There are also soft models as have been described by Wold (1991). Regression based Profit and Cost models have been developed by Shamim et al. (2017a). Chaudhary et al. (1996) has developed a financial accounting style method for fiscal deficit and debt scenario analysis. Lian et al (1996) considered twenty different factors in order to assess the feasibility of any country for the sake of investment amongst which thirteen are market related factors and seven are country related factors. A study at the International Finance Corporation listed institutional uncertainties related to private investment (Brunetti et al.; 1997). Anjum (2003) applied classification methods for the understanding of the root causes of twin banking and currency crisis of 1998 in Asian markets (i.e., Asian financial crisis). Other applications of classification techniques are in credit scoring and evaluation, bankruptcy predictions, insurance underwritings, fraud detections, financial performance prediction, bond rating analysis, credit risk assessment, to forecast daily changes in seven financial stocks' prices and other applications in finance (Anjum 2013).

Before mining the data, we decided on the training and the testing data. Cross-validation is gaining ascendance and is probably the evaluation method of choice in most practical, limited data situations. Some situations involve predicting class's probabilities rather than classes themselves, while others involve predicting numeric rather than nominal values. The basic principle for accuracy measurement uses an independent test set rather than the training set for performance evaluation, the hold-out method, cross validation apply equally well to numeric prediction. However, the basic quality measured offer by the error rate is no longer appropriate. Errors are not simply present or absent, they come in different sizes. In some cases, the test data might be distinct in nature from the training data. If the data sets were amalgamated before training, performance on future data in different circumstances may not give good results. The next step after the error was estimated, the test data was bounded back into the training data to produce a new prediction value for new data. The more data used in training the more accurate prediction value is gained. More importantly, we had to deal with is to make sure that both the training and the testing data sets represent and contain the most of the classes held by the attributes. The techniques are usually done by using the stratification procedure.

Diagram 1. JCLW's Knowledge Flow Model: used in our analysis



Different learning methods for various data mining algorithms have been introduced e.g. supervised learning, unsupervised learning, semi supervised learning, reinforcement learning, transduction learning and learning to learn etc. (Ahmed and Zeeshan 2014) and Hall et al. (2010)). Various different data mining techniques and their description have been provided in Table 1. Our analysis has focused on nine attribute selection algorithms from Java Class Library Weka (JCLW) along with their variety of alternate features and search methods, which has totaled our search algorithms to thirty-two.

The nine algorithms, which have been used here, are Correlation AttributeEval, GainRatio AttributeEval, Cfs SubsetEval, InfoRatio AttributeEval, OneR AttributeEval, SymmetricalUncert AttributeEval, ReliefF AttributeEval, Wrapper SubsetEval and Principal Components. Moreover, in order to automate various workflows, JCLW's Knowledge Flow application may be used. JCLW's Knowledge Flow's sample application has been shown in diagram 1. We have run thirty two algorithms in JCLW for either 10-cross validation or on full training set if validation option was not available each time a limited set of indicators were selected along with different and varying values for the selected criteria.

IV. RESULTS AND ANALYSIS

4.1. Results

The present study used various attribute selection algorithms available in JCLW in order to get the most relevant indicators, out of total 188, for the prediction of our target variable (or class) which is FTSE BURSA Malaysia EMAS Sharia

price index (FTBMEMA) using numerical class as well as nominal class. Nominal classes were divided into various bands (zones), based on the standard deviation based distance from mean criteria. Names of various zones of nominal class were named as Jackpot, Bingo, Average, Danger Zone and Risky Red bands. The selected attributes by various big data algorithms from data mining, was then ranked based on various characteristics i.e., the order of their rank by algorithm and the statistical significance measures. The techniques for ranking like analytical hierarchal process can be found in Anjum (2014d & 2015) which have been applied in various context including for the selection of Islamic based risk as have been described in Anjum (2016). We have run thirty two versions of nine algorithms in JCLW for either 10-cross validation or on full training set if validation option was not available in the algorithm. This exercise has selected 18 variables from 188 variables, which is a significant reduction in the number of variables to be used for any prediction methodology including regression analysis.

There are different types of statistical measures to judge the significance and accuracy of empirical models. The choice of the statistical measures depends on the relevance of the model at hand. Some of descriptive statistics and commonly used statistical measures have been applied in Shamim et al. (2013, 2015 & 2017b) and empirical studies regarding importance of various predictive classifiers reveal that there is no objective conclusion about superiority of one classifier over the other, rather performance of any classifier depends on the nature of problem, type of dataset to be used and behavior of variables (Anjum 2014b). Besides the use of different models, the interesting thing is to know how predictive a model is? One of the interesting fact, which has become evident from this study, is that not all of the spot values of our primary variables have proven to be useful for the prediction of our target variable named FTSE BURSA Malaysia EMAS Sharia price index (FTBMEMA). The spot values of FTSE BURSA Malaysia Hijrah Sharia Price Index, FTSE BURSA Malaysia EMAS Price Index, Kuala Lumpur SE Industrial Price Index, GMC Strait Trading Index of Singapore (GSTI-SNG), Kuala Lumpur SE Industrial Production Price Index (KLSEIP-PI), Kuala Lumpur SE Technology Price Index, FTSE BURSA Malaysia Mid-70 Price Index (FBMM70-PI) and Singapore$-US$ exchange rate variables have been found as relevant as per attribute selection algorithms. Other selected variables are five days lag values of KLSEIP-PI, FBMM70-PI and FTSE Malaysia Consumer Goods Price Index (FTMCGL-PI), 10 days lag values of Kuala Lumpur SE Technology Price Index and FTMCGL-PI, twenty days lag values of GSTI-SNG close, thirty days lag values of GSTI-SNG close and Kuala Lumpur SE Consumer Production Price Index as well as fifty days lag values of Gold price. The attributes of rest of the primary and derived variables have not proven to be statistically significant through the out-of-sample validation of 188 of total variables included in the analysis. One of the advantages of the data mining algorithms over the other methodologies in the literature is that data mining algorithms give the in-sample probabilities explicitly in their evaluation measures. Besides, cross-fold validation criteria has been used for establishing the statistical significance of the results. In order to see the effectiveness of our attribute selection success for prediction purposes, we have used the selected attributes by using regressions analysis and the results have been show in Table 2. All, but three of the eighteen selected variables which have been

excluded from the regression results in Table 2, have worked well in the regression analysis with statistically significant contributions.

Table 2. Results obtained from Linear Regression

List of Recommended Indicators Used in Regression	Standardized Coefficients			Correlations	
	Beta	t	Sig	Zero Order	Partial
GMC-STI-SNG-SPOT-CLOSE	.083	6.594	.000	.610	.132
SNG-$-XR-SPOT-	-.041	-2.741	.006	-.603	-.055
FTBM-MID70-SPOT-M$	.093	2.973	.003	.975	.060
FTBM-EMAS-SPOT-M$	.360	8.345	.000	.981	.167
KLSE-INProd-SPOT-M$	.129	8.516	.000	.961	.170
KLSE-IND-SPOT-M$	-.108	-7.516	.000	.940	-.151
KLSE-TEC-SPOT-M$	.060	2.951	.003	.079	.060
FTBM-HjrhShrh-SPOT-M$	.397	26.248	.000	.977	.470
10D-MA-FTM-ACGL-CPI-M$	.043	5.464	.000	.786	.110
10D-MA-KLSE-TEC-M$	-.099	-4.998	.000	.064	-.101
20D-MA-GMC-STI-SNG-CLOSE	-.435	-11.341	.000	.623	-.224
30D-MA-GMC-STI-SNG-CLOSE	.442	13.468	.000	.631	.263
30D-MA-SNG-$-XR	.198	11.923	.000	-.600	.235
30D-MA-KLSE-ConProd-M$	.133	9.826	.000	.926	.195
50D-MA-GldSptPI-CLOSE	.024	3.363	.001	.476	.068

Notes: Constant non-standardized coefficient value is -5046.4 with t-value of -12.481 (i.e. significance at 0.00)

4.2. Analysis

The study used thirty two versions, of nine algorithms, applied to 2550 observations of 188 variable and only 18 variables were short listed in the process, which was relevant for the prediction of the various zones of FTSE BURSA Malaysia EMAS Sharia price index (FTBMEMA). This significant reduction in the number of variables for use in any prediction methodology including regression analysis is a great achievement of the proposed methodology. The results obtained in this analysis are of significant nature and thus boosted the importance of the attribute selection recommender system, which have been discussed marginally here, and with details in Anjum (2017c). The Adjusted R Square for the regression model, achieved using data mining based selected variables is 0.986 with 15 degrees of freedom. The ANOVA results show that F-stat is significant at 99% level. All t-values for all the variables are significant at over 95% level of significance. The correlations are, however, high among the variables and multicollinearity may be an issue in this regression, which necessitates the application of methods to manage the issues. The study did not extend its scope to that extent, as the primary purpose of this study was to test the applicability of data mining based reduction of variables algorithms and its successful application to achieve regression results with valid Adjusted R Square and t-values. The study has shown that for any prediction problem, it is always a wise strategy to use attribute selection algorithms

before using the dataset in prediction, classification or forecasting algorithms or in regressions because of its demonstrated potential in saving time and efforts spent which may have otherwise been wasted either on plotting hundreds of variables against each other or on doing unnecessary juggling with running regressions in a traditional trial and error fashion in order to arrive at the best selection of variables and/or results. All this in turn reduces the excuses for regressioneering (Anjum 2014a) or regressgineering (Anjum; 2014c) and thus makes the completely analytical or experimental exercise more scientific as well as legitimate one.

The results obtained in this study are interesting ones and have shown that all variables, except four, have positive relationship with target variable. The four variables which have negative coefficients in regards to our target variable are Singapore$-US$ exchange rate, Kuala Lumpur SE Industrial Price Index, ten days lag values of Kuala Lumpur SE Technology Price Index and twenty days lag values of GSTI-SNG close. Interestingly the spot and thirty days lag values of GSTI-SNG close have positive relationship with target variable. Some results of this study are important discoveries as well. The results are where the popular theoretical perceptions of the Islamic finance literature have been contradicted for like the findings of negative coefficients for three independent variables i.e., Kuala Lumpur SE Industrial Price Index, ten days lag values of Kuala Lumpur SE Technology Price Index and twenty days lag values of GMC Strait Trading Index of Singapore close. Sharia index, being comprised of companies which are Sharia compliant, meaning that the companies are supposed to avoid Gharry (uncertainty) through the reflection of the real activity in the real economy into the index, should have positive relationship with the indexes of real activity sectors like industrial and technology. However, it did not seem the case in case of FTSE BURSA Malaysia EMAS Sharia price index (FTBMEMA). Moreover, the relationship of FTBMEMA with GMC Strait Trading Index of Singapore (GSTI-SNG) was cyclical. This is evident from the fact that three series selected by the data mining algorithms for GMC Strait Trading Index of Singapore (GSTI-SNG) variable, i.e., spot values, twenty days lag and thirty days lag values and while coefficients for spot and thirty days lag values of GSTI-SNG are positive with the values of 0.083 and 0.442 respectively, the coefficient of twenty days lag values of the variable is negative (-0.435).

V. CONCLUSION AND RECOMMENDATION

5.1. Conclusion

This study has attempted to find out an efficient way, using data mining algorithms, of finding statistically relevant explanatory factors for the estimation of regression where FTSE BURSA Malaysia EMAS Sharia price index (FTBMEMA) is dependent variable. The study has used 23 relevant indexes and two exchange rates as primary variables. Momentum indicators for short, medium and long term for the 25 variables have also been calculated for the variables, which has totaled to 188 original and derived variables combined. The problem of reducing the number of variables, from 188 to reasonable size with statistically significant characteristics, for regression estimation as well as other predictive models like neural networks etc. is not only significantly important but also difficult one.

In this preliminary study, data mining attribute selection algorithms along with cross validation criteria were used, through the use of Java class library Weka (JCLW), for reducing the number to statistically relevant variables for our regression estimation in an effort to forecast various performance parameters for FTBMEMA like performing either in a mean performance range, having jackpots and bingos or falling into danger zones. The importance of reducing the variables from 188 to fifteen is a significant achievement of this approach of using nine attribute selection models of JCLW's which has totaled to thirty two when nine basic algorithms have been run choosing their diverse features and search methods. The results obtained in this study are interesting in the sense that all, except four variables, have positive relationship with target variable. The four variables which have negative coefficients with target variable are Singapore$-US$ exchange rate and Kuala Lumpur SE Industrial Price Index as well as short run momentum indicator of Kuala Lumpur SE Technology Price Index and twenty days lag values of GSTI-SNG close.

5.2. Recommendation

As far as the thoughts on further improvements are concerned in future research on the same lines as well as recommendations for academicians, the application of other algorithms in data mining and techniques of deep learning like neural networks and genetic learners for the generation, may be applied for further enhancement in prediction quality and accuracy comparisons may be based on the selected attributes. We are currently applying, as well as aiming to enhance this study further on similar lines in another project. Besides, an easy but sophisticated value addition is on cards regarding weighting and aggregation techniques as well. Moreover, as the current data has comprised of a period which fall before, during and after sub-mortgage crisis of 2008 and financial meltdown of 2009, it may make more sense to do analysis and experiments based on separate datasets for the periods of before, coincidence and after sub-prime mortgage crisis of 2008 and also to do the cross period training and testing to get deeper insights on various algorithms.

Recommendations for practitioners are that portfolio managers may hedge and/or rebalance through shredding off or putting on their portfolios to increase their returns by properly managing cyclical relationships between various variables, which have been discussed in this study. Regulators in various industry sectors of the financial market, i.e., commodity sector regulators, stock market regulators, regulators in precious commodity sector and central banks as regulators of foreign exchange may enhance their understanding of the effects of cyclical nature of inter-relationships of various variables from various closely connected sectors or market segments and therefore device their strategic vision and policies based on the facts of cyclical influences within an echo system of closely related sectors.

REFERENCES

Ahmed, Z. and Zeeshan, S. (2014). Applying WEKA towards Machine Learning with Genetic Algorithm and Back-propagation Neural Networks. *Journal of Data Mining Genomics Proteomics*, *1(2)*, 157-160.

Anjum, S. (2003). Early warning system for financial crisis: a critical review and application of data mining approach. Japan: GSID, Nagoya University. Ph. D. Dissertation.

Anjum, S. (2013). Algorithms for Predictive Classification in Data Mining and Evaluation Methodologies. *Journal of Industrial and Intelligent Information* (JIII), *1(2)*. June.

Anjum, S. (2014a). Statistical Software a Regression Diagnostic Reporting with Fuzzy-AHP Intelligent Zax. *Lecture Notes in Software Engineering*, *2(1)*. February.

Anjum, S. (2014b). Composite Indicators for Data Mining: A New Framework for Assessment of Prediction Classifiers. *Journal of Economics, Business and Management (JOEBM), 2(1)*.

Anjum, S. (2014c). Systematic Risk Outliers and Beta Reliability in Emerging Economies: Estimation-Risk Reduction with AZAM Regression. *Review of Integrative Business and Economics Research (RIBER),* 3(1).

Anjum, S. (2014d). Quantification of Fiduciary Risks: Islamic Sources of Funds, Neo-Institutionalism and SARWAR Bank. *Journal of Islamic Banking and Finance*, *2(1)*.

Anjum, S. (2015). Market Orientation, Balance Sheets and Risk Profile of Islamic Banks. *International Journal of Economic Policy in Emerging Economies*, *8(4)*.

Anjum, S. (2016). Banking Automation with Sustainable Hedging for Information Risks: BASHIR Framework for Clouds Computing. *Advanced Science Letters, 23(11),* 11609-11612(4). USA: American Scientific Press.

Anjum, S. (2017a). *Risk Magnification Framework for Clouds Computing Architects in Business Intelligence*. Proceeding of International Conference in Information Education and Technology (ICIET 2017). The Association of Computing Machinery (ACM).

Anjum, S and M. Kamaluddin (2017b). Country Risks in Selected World Economies: Application of Niche Methodology. *Review of Integrative Business and Economics Research*, 6(4).

Anjum, S and Shamim, F. (2017c). Shariah Stock Index Jackpots and Red Zones: Big Data Algorithms to Recommender System. 8th International Conference on Islamic Banking & Finance: Risk Management, Regulation, and Supervision (8th ICIBF). Sultan Qaboos University. Oman. Dec.

Avery, C. N., Chevalier, J. A. and Zeckhauser, R. J. (2015). The CAPS Prediction System and Stock Market Returns. *Review of Finance*, *1(29)*.

Bouckaert, Remco R., Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald & David Scuse. (2016). *WEKA Manual for Version 3-8-1*. New Zealand: University of Waikato. Hamilton.

Breiman, L. (2001). Statistical modeling: two cultures. Statistical Science. *Institute of Mathematical Statistics*, *16(3)*, 199-215.

Brunetti, A and Weder, B. (1997). Investment and Institutional Uncertainty: A Comparative Study of Different Uncertainty Measures. *International Finance Corporation (IFC) Technical Paper*, 4. The World Bank, Washington D.C.

Chaudhary, M. A. and Anjum, S. (1996). Macroeconomic Policies and Management of Debt, Deficit, and Inflation in Pakistan. *Pakistan Development Review, 35(4)*. Part II. Winter.

Edward R.D and Magee J. (2001). *Technical Analysis of Stock Trends*. 8th ed. Publisher: St. Lucie Press.

Hall, M., Eibe F., Geoffrey H., Bernhard P, Peter R. & Ian H. W. (2010). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11(1)*, 10-18.

Jones, C.P. (2007). *Investments. 14th ed*., Publisher: John Wiley & Sons.

Kovalerchuk, B and Vityaev, E. (2000). Data Mining in Finance: Advances in Relational and Hybrid Methods. USA: Kluwer Academic Publishers.

Langdell, S. (2002). Examples of the use of Data Mining in Financial Applications. *Financial Engineering News*, 25. April.

Radelet, S and Jeffrey D. S. (1998). The East Asian Financial Crisis: Diagnosis, Remedies, Prospects. *Brooking Papers on Economic Activity, 1*, 1-90.

Reilly, F. K. and Brown, K. C. (2012). *Investment Analysis and Portfolio Management. 10th edition*. USA: Cengage Learning.

Rockefeller, B. (2004). *Technical Analysis for Dummies*. Publisher: Wiley.

Shamim, F. and Anjum, S. (2013). Technology Diffusion in the Japanese Finance Industry: An Exploration. *International Research Journal of Applied Finance, 4(12)*.

Shamim, F., Anjum, S and A. A. Wakil. (2015). Banking Risk and Operating Efficiency Measures in the Era of IT. *Accounting and Finance Research, 4(1)*. Sciedu. Canada.

Shamim, F., Nobuyoshi Y and Anjum, S. (2017a). Clicks Business of Deposit Taking Institutions: An Efficiency Analysis. *Journal of Economic Studies, 44(6)*. Emerald Publishing. Thomson Reuters.

Shamim, F. and S. Anjum (2017b). Economic and Financial Agents on Islamic Finance. submitted to Thunderbird International Business Review. John Wiley & Sons. USA

Tjung, L. C. O., Kwon, K. C. Tseng, and J. Bradley (2010). Forecasting financial stocks using data mining. Craig School of Business. CSU. Fresno

Wold, H. (1991). Soft Modeling: The Basic Design and some Extensions. In essays in honour of Karl A. Fox, edited by Tej K. Kaul and Jati K. Sengupta. Elsevier Science Publishers.

Yadav, A. K, Hasmat M. & S. S. Chandel. (2014). Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction model. *Renewable and Sustainable Energy Reviews, 31*, 509–519.