JOURNAL
OF
APPLIED
QUANTITATIVE
METHODS

PREDICTING SHARIA STOCK PERFORMANCE IN INDONESIA STOCK MARKET USING SUPPORT VECTOR MACHINE ALGORITHM FOR IMBALANCE DATA

Retno MAHARESI[1]
Department of Information Technology, Gunadarma University, West Java, Indonesia

E-mail: rmaharesi@staff.gunadarma.ac.id

Sri HERMAWATI
Department of Management, Gunadarma University, West Java, Indonesia

E-mail: srihermawati@staff.gunadarma.ac.id

ABSTRACT:

The paper discusses the practical implementations of using the support vector machine (SVM) algorithm for imbalance data to predict the stock performances in the Indonesian stock market. SVM algorithm for imbalance data was used to model various financial ratios as independent variables to investigate indicators that significantly affect the stock's performance of large market capitalization companies which were actively traded over the last three-year periods. The model selections, namely the imbalance and the balance SVM model with dummy variables representing the appropriate weights were carried out using 10-fold cross validation methods integrated with a grid search procedure for parameter optimization. The study identified and examined six financial ratios commonly used by the stock analysts without considering macro economic variables was able to classify the performances of the companies into two categories "good" or "poor" based on the prices proportion of two consecutive periods. The result suggested that the proposed method yield competitive performance in terms of prediction accuracy level as compared with its competitors.

Key words: *Classification, financial ratios, prediction, stock's performance, support vector.*

1. INTRODUCTION

Predicting stock performance is a very difficult as well as a challenging problem. In the history of stock performance forecasting literature, no comprehensive or accurate model has been recommended to date. Hence the usage of relevant financial information to make good investment decisions in the stock market is important. Having made careful observations for the interdependence among some relevant variables such as interest rates, bond returns and volatility in affecting the probabilities of different types of stock market crashes, stock analyst should as well pay attention to emerging stock markets, since a certain local crash can be seen as a bad signal in other market places.

A stock's performance can, to some extent, be analyzed based on financial indicators presented in the company's annual report. In which, the annual report provides a vast amount of information that can be use to produce various financial ratios. Many literatures told that financial ratios can be used for assessing future stock performance namely to

project future stock price trends based on those previous values. Ratio analysis therefore becomes a key parameters used by stock analyst to determine the intrinsic value of stock shares. The study of financial ratios emerged after stock market crashes in the 1998s and 2009s in Indonesia. Today, ratios are also used in fundamental analysis to predict future company's performance. As the need in this type of analysis to grow, various new ratios, such as book value and price/cash earnings per share, have been included for share valuation. The level of importance given to these ratios differs from industry to industry and from one country to another. Thus, selecting appropriate ratios is very crucial in increasing the prediction success rate.

The objective of this paper is to analyze financial data in order to develop a simplified model for interpretation. This paper intents to build a model for classifying sharia stocks belong to *Indonesian Syariah Stock Index* (ISSI) into two categories (good or poor), based on their rate of return. Here a company's stock is classified as "good" if its share prices to be greater than of the previous period's price, provided that the price ratio of the two consecutives periods is above the predefined value, e.g., 130 % in the data sample. In this study, the SVM algorithm has been used to classify selected companies, based on their performance. The SVM algorithm is used to predict the categorical value (poor or good) of stock performances by classifying the data of feature variables. It involves optimizing the margin value to locate the hyper plane in between the two different classes of good and poor stocks. This is carried out by first transforming the feature variables using nonlinear implicit kernel mapping into a possibly higher feature dimensions to make the data to be able being delineated by the hyper plane.

Among all of the stocks listed in the Indonesia Stock Market, not all of them be grouped into sharia index. A certain stock considered to be a sharia stock after an examination and investigation processes carried out by the board of Islamic sharia commission, namely *Dewan Sharia Indonesia* (DSI). The board of commission reports the investigation results to the Government body namely Bapepam-LK. To enable helping investors for doing investment on these shares, Bapepam-LK announced a list of stocks considered to follow sharia principle called *Daftar Efek Sharia* (DES) every semester. DES contains a set of stocks in the stock market which do not violate the sharia principles. Generally the shares included in DES involve not only stock issued by public companies, but also other kinds of shares such as mutual funds, bond, obligation and some other sharia effects. The bench mark of market performance for all Indonesia sharia stocks (ISSI) is 30 most actively transacted stocks during the last year of examination process carried out by DSN, called Jakarta Islamic Index (JII). Criteria to decide whether a certain stock to be in the JII list are: The main business of share issuing company does not violating the sharia laws and already listed for more than three months (except it is included in the top 10 big market capitalization). The stocks should fulfill the following conditions: For which is based on its annual or semester financial report, it must have liability ratio upon maximal asset to be around 90%, to be considered in the 60 top stocks ranked based upon biggest market capitalization averages during the last year, be considered in the 30 stocks ranked based upon average of regular trading liquidity values during the last year.

2. REVIEW OF LITERATURES

In the last 10 recent years, there has been a greater focus of the market because of

its rapid growth and its increasing potential for global investors to the Indonesia stock market. Emerging-market returns are usually more predictable than of the developed market returns because it is more likely to be highly influenced by local information than those of the developed markets (Harvey, 1995). Several written works examined the cross-sectional relationship between fundamental variables with indicators such as earnings yield, cash flow yield, book-to-market ratio and others to predict stock returns in the case of developed markets as well as the use of financial ratios for classifying the performance of firms (Kato et al.1996; Jung & Boyd, 1996). These indicated that a good prediction performance can be obtained based on the financial ratios data.

In light of the market's growing importance, more attention has been directed to studies concerning different classification techniques for measuring stock performance. One of the technique is Logistic regression (LR), that can be used for predicting the presence or absence of a characteristic or outcome based on values from a set of predictor variables (Lee, 2004). The model can be used for classifying firms as either defaulters or non-defaulters. The model was focused upon the ability to rank the defaulted and non-defaulted firms, based upon the failure probability to be more favorable. In order to apply logistic regression, one must consider two kinds of error rate, namely type I and type II rates in the selection of the optimal cut-off probability. Hence, there is a subjectivity of the choice of these misclassification costs in practice, which is the weakness of the LR procedure (Zavgren, 1985).

Some studies generally reported that data-mining techniques such as artificial neural networks (ANN) and support vector machine (SVM) were better suited to detect stock-price movement, in terms of classification accuracy than multivariate statistical techniques such as discriminant analysis or LR. Among them were (Cheng 1996, Van & Robert 1997, Öğüt 2009) who used data mining techniques for modeling financial time series. Min & Lee (2004) showed that SVM outperformed the ANN for predicting business failure in Korea for their comparative study. Min & Jeong (2009) compared prediction accuracy of a binary classification model with other methods such as multi-discriminant analysis, LR and ANN, who showed that their model can be a promising alternative to consider among the existing model forbankruptcy prediction. A comprehensive review of various work related to bankruptcy prediction problems carried out by Kumar & Ravi (2007) found that neural network is the most widely used technique. Mostafa (2010) showed that neuro-computational models, called quasi-Newton training algorithm is useful tools in forecasting stock exchange movements in emerging markets. The training algorithm they used produces fewer forecasting errors, compared with other training algorithms. This is because the robustness and flexibility of training algorithms serve neuro-computational models to be expected to outperform traditional time series techniques such as regression and ARIMA in forecasting price movements in stock markets.(Guresen et al. 2011) evaluated the effectiveness of neural network class models, namely multi-layer perceptron (MLP), the dynamic artificial neural network and hybrid neural networks that used generalized auto-regressive conditional heteroscedasticity (GARCH) to extract new input variables in stock market predictions. (Li et al. 2010) used a 30 times hold- out method to build a better model for predicting stock returns, along with the two commonly used methods in data mining algorithms (SVM and *k* Nearest Neighbor) are used. They concluded that the homogeneous multiple classifiers utilizing neural networks by majority voting perform best to predict stock returns. (Swiderski et al. 2012) demonstrated an approach for an automatic assessment of the company financial condition and developed the computerized classification system, applying a

certain representation of data and then using support vector machine (SVM) as the final binary classifier. The application of this method to the firm's financial condition assessment has shown to be superior with respect to the classical approaches.

2.1 SVM algorithm

SVM model for imbalance data is similar to SVM model of balance data, but is qualified in models where the dichotomous dependent variable shows that the number of certain outcome is very much bigger than of the other. SVM model for imbalance data can be used to predict the membership of each unit sample based on the independent variables or features values in the model. The SVM algorithm yields coefficients for each independent variable based on model selection for optimal parameter using a data sample. Support vector machine models (SVM) which contain nonlinear and noise components are widely used in practice (Hsu et al., 2013). The parameters of an SVM model are commonly estimated by learning from the data sample.

The particular advantage of SVM is that, through the application of implicit nonlinear mapping, i.e. by implementing a kernel function; for mapping the feature space of independent variables into higher dimensions. This likely to produce a new set of data sample in the higher dimensions of features space and then making possible for the usual hyper plane to delineate in between the two different groups of sample (Lee, 2004). The predictor values from the analysis (*-1* or *1* outcome) can be interpreted as a membership in the target groups (categorical dependent variables).

Existing literature indicates that SVM model for imbalance data has been rarely used to build a model for predicting out-performing sharia shares. SVM model has been used mostly for predicting financial distress and business failure. It has not been used for predicting sharia share performance in Indonesia. In terms of stock market investment destination, Indonesia is a good performing emerging market. In this context, the present study will provide useful method of stock analysis to shareholders and potential investors to help them to make good decisions regarding the investments.

3. RESEARCH OBJECTIVE AND METHODOLOGY

In this study, the relation between financial ratios and stock performance of the firms has been examined using binary Support Vector Machine algorithm for imbalance data. The earlier studies cited above have generally shown support vector machine algorithm for imbalance data, as used in the finance field can be an effective tool for decision makers. It has also been recognized fundamentally that financial ratios can noticeably enhance the forecasting model ability of stock price. The objective of this study is to build a model using financial ratios of the firms for predicting out-performing shares in Indonesia Sharia Stock Index (ISSI). The goals of this study are to answer the questions: Can the stock's prices proportion of two consecutive periods be explained by using financial ratios? (2) Can the stock's returns through the use of stock's prices proportion of two consecutive periods be analyzed using a support vector machine model for imbalance data?

3.1. Analysis of SVM model for imbalance data

The classification task in SVM model is performed by separating hyper plane for which every unit data in the sample can be assigned to one of the two different classes

containing in the data sample. The hyper plane produced by the SVM algorithm is formed such that the distance between a subset of data sample, known as vector supports should be maximal towards the hyper plane in order to have a better decision for learning process among the set of *n* data sample of *p* dimensions feature variables. The relationships between independent variables which contains several financial ratios and the dependent variables, in this case is stock categories of 'good' or 'poor' stock obtained from the learning process is then used for predictions. The outcome of SVM algorithm is a function that gives the decision of classification (good or poor) on the substitution of independent variable into the decision function Hofmann et.al (2008) obtained in the learning or training process. Hence SVM models allow one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete or a mix of any of these. The relationship between the hyper plane *f(x)* and the support vectors data are described by Figure (1). Here both the vertical and horizontal axis denoted the explanatory variables x_1 and x_2.

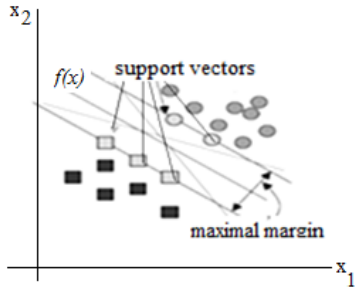

Figure 1.Hyper plane obtained by SVM algorithm in R^2 space

Real data rarely show linear separable phenomenon, hence a more suitable model to tackle for that data should accounts for non-linear as well as noise components. Based on a set of input quantities $\mathbf{x}_i$, $y_i \in \{-1, 1\}$; C, $K(\mathbf{x}_i, \mathbf{x}_j)$ for $i, j = 1, \ldots, n$ where $\mathbf{x}_i$ is a vector of *p* feature variables of a data sample. The assigned classification for x_i is denoted by using binary label {-1, +1}, with C is a set up parameter to represent the penalty cost of noise presence ε_i in the data. The kernel function $K(x_i, x_j)$with parameter γ represents an implicit nonlinear mapping in the data feature spaces, namely $\varphi: R^p \rightarrow R^m$ where $m \geq p$. The original SVM model formulation due to (Cortes & Vapnik 1995) for data sample of size *n* is

$$min \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i \qquad (1)$$

Subject to:

$$1 - y_i\left(\langle\varphi(\boldsymbol{x_i}), \varphi(\boldsymbol{x_j})\rangle + b\right) - \varepsilon_i \leq 0$$

$$\varepsilon_i \geq 0 \quad for\ i = 1, 2, \ldots, n.$$

The solution of the SVM optimization problem obtained by first to get the solution for the dual optimization problem in terms of Lagrange multiplier α_i:

$$max \quad L(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j \sum_{j=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_j) \qquad (2)$$

subject to: $0 \leq \alpha_i \leq C,\ 0 \leq \beta_i \leq C,\ C - \alpha_i - \beta_i = 0,\ \sum_{i=1}^{n}\alpha_i y_i = 0$ for $i = 1, \ldots, n$.

A Quadratic Programming software (QP) may be used to get this solution. From the obtained optimal solution one arrives to the decision function after a substitution for optimal α_i into the expression of model's parameter (***w***, *b*). The prediction classification for the whole data is obtained based upon these parameter values. Here the optimal parameters values are obtained using grid search scheme implemented in the cross validation method based on the data training. After obtaining the optimal parameter values over the grid search areas then the prediction equation as shown in Equation (3) can be obtained.

$$y\,(\boldsymbol{x}_{test}) = sign(\textstyle\sum_{i=1}^{n} \alpha_i y_i\, K(\boldsymbol{x}_i, \boldsymbol{x}_{test}) + \; b) \qquad (3)$$

The above SVM model works well on a set of balance data. However, for the case of imbalance data, where the number of 'good' and 'poor' stocks differ significantly, a more suitable model, namely SVM model for imbalance data should be used to obtain a more accurate model (Anand et al, 2010; Akbani et al, 2004; Hur and Weston, 2008). For the case of imbalance data, there are several other methods to make the imbalance data to be more balance. The most common methods applied for imbalance data include changing size of training datasets, cost-sensitive classifier (Akbani et al. 2004, Han et al. 2005, Chawla et al. 2002). Several other methods to deal with imbalance data had been as well proposed, which include modified SVMs, k nearest neighbor (kNN), neural networks, genetic programming, rough set based algorithms, probabilistic decision tree and learning methods (Ganganwar, 2012). Those methods were reported to have good classification performances on the imbalance data sets. To note for method of resizing data which include oversample the minority class or under sample the majority class can cause to non randomness of sample distribution and hence its distribution no longer approximates the target distribution. While with under sampling, this might discard instances with valuable information. For this reason, here we come out to the addition of two dummies variables in the feature variables by implementing a weighting scheme for each data sample according to the class assignment of each data sample in the training set. Having performing this stage, the usual SVM algorithm can be implemented to obtain the decision function for classification.

Formulation for *C* parameter for the case of imbalance data needs some modification according to the following argument: The Probability to correctly assigning a unit sample to the right class i.e., positive or negative classes can be estimated using the following conditional probability formulae: $P(True) = P(True \cap pos) + P(True \cap neg) = P(True|pos)P(pos) + P(True|neg)P(neg)$ or $w_{pos}P(True|pos) + w_{neg}P(True|neg)$.

The SVM Model in (1) used the penalty cost *C* equally to all data classes, which means to assign equal weight of ½ for the two different classes. As a result, the decision function tends to classify every unit sample to the majority class (Hur & Weston, 2008). To overcome this bias, the weight assignment needs to be changed proportionally according to the number of positive and negative membership. If for example the number of negative data (n^-) is much greater, then the probability of negative misclassification becomes small, therefore, the penalty cost on the negative class (C) should be a small positive number. On the contrary, if for example, the number of data to have positive sign (n^+) is very small, then the probability to occur misclassification for the positive class increases, hence the penalty cost for the positive class namely, C_+ should be a big positive number, hence equation $C_+ n_+ = C_- n_-$ is obtained or

$$C_+ = \frac{n_-}{n_+} C_- \quad \text{and} \quad C_- = 1\, C_- \,. \qquad (4)$$

Applying the weight notation as before to yield: $w_{pos} = n_-/n_+$ and $w_{neg} = 1$. It is preferable to use only a penalty cost parameter C_- instead of both, this is done by substituting (5) for the total penalty cost in the second term of the objective function in (1), to yield

$$\sum_{i:=I}^{n_-} \varepsilon_i = \sum_{i:\in I+}^{n_-} \varepsilon_i \; C_+ + \sum_{i:\in I-}^{n_-} \varepsilon_i \; C_- \; . \tag{5}$$

By then the objective function for the case of imbalance data is given by (6)

$$min \frac{1}{2}\|w\|^2 + C_- \left(\sum_{i \in I+}^{n_+} \frac{n_-}{n_+} \varepsilon_i + \sum_{j \in I-}^{n_-} \varepsilon_i \right) \tag{6}$$

Based on this result, the primal form of Lagrange Equation can be obtained in the same way as before and the dual optimization problem for imbalance SVM model can be solved for the same Lagrangian multiplier α_i imposed by addition for dual constraints as follows.

$$0 \le \alpha_i \le C_- \text{ if } y_i = +1 \text{ and } 0 \le \alpha_i \le C_+ \; if \; y_i = -1. \tag{7}$$

Due to the α_i values for data in positive class which are much bigger than those of the negative class, then this enable the hyper plane to be pushed closer to the minority class boundary. Hence the improvement for the generalization ability can be expected. Following (Luts et al. 2010) (6) can be written in terms of dummy variable v_i:

$$\frac{1}{2}\|w\|^2 + C_- \sum_{i=1}^{n} v_i \, \varepsilon_i \text{ , where } \quad v_i = \begin{cases} \frac{n_-}{n_+} = w_{pos} & jika \; y_i = +1 \\ 1 = w_{neg} & jika \; y_i = -1 \end{cases} \tag{8}$$

From here, based on Equation (8), addition of two feature variables presented as two dummy variables can be obtained. According to the class assignments of each unit sample, a data sample comes from positive class have the value equals to n_-/n_+ in the first dummy variable, w_{D1}. While the data sample belong to the negative class will be assigned value equals to *1* in the second dummy variable, w_{D2}. As a result, the decision function contains two more variables. For the final model which is usually intended to make a forecast for the newly known data, $\boldsymbol{x}_{new} = (x_1, x_2, \ldots, x_p)$ the dummy variables above are not yet known, this problem is overcome using a pessimist (or optimistic decision). For pessimist decision, the second dummy variable x_{D2} is assumed to be true and the first dummy variable is false.

In summary, based upon a sample of features data presented in vector $\boldsymbol{x}_i$ and the corresponding labeled value of classification y_i, written in the matrix form $X = \{\boldsymbol{x}_i, y_i\}$ and a radial basis kernel function *K*, the hyper plane function for classification task based on SVM algorithm is obtained by getting the optimal solution for the dual variables α_i of optimization problem (6), to yield the classification decision function as given by Equation (3). Therefore the prediction function obtained from the model selection process has the following formulation:

$$y\,(\boldsymbol{x}_{new}) = sign\left(\sum_{i=1}^{p} w_i \,.\, x_{new} + \; b \; + w_{D2}.x_{D2}\right). \tag{9}$$

Where b is the constant parameter of the hyper-plane, w_i is the estimates of the vector coefficient of independent variables and w_{D2} is the coefficient estimate of the second dummy variable.

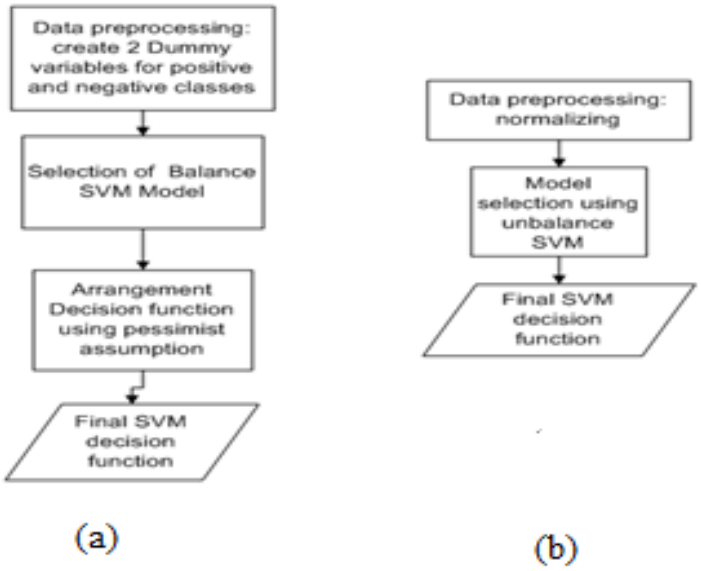

Figure 2. Flow process diagram for imbalance data SVM model selections.

3.2. Application SVM algorithm for imbalance data

The financial data used throughout this analysis was collected from the Web link www.jsx.co.id. The sample for the study was drawn from companies whose shares are most actively traded which belong to the Jakarta Islamic Index. A purposive sample taken from three years periods (2010-2012) of annual financial reports was selected for classification purposes. The data set contain $n = 142$ most active Sharia stocks during year of 2012 in Indonesia Stock Market. Based on these Sharia stocks the data from the previous years of 2011 and 2010 were also included. All the stocks data for all periods must have the six independent variable values. The independent variables (feature variables) are based on indicator usually used by practitioners in stock market analysis, those are ROE (return on equity), DAR (debt to asset ratio), EPS rate of growth, PER (market price to earnings ratio), PEG ratio (price to equity ratio divided by EPS) and NPM (net profit margin). For the purpose of carrying out SVM data modeling, first to do is acquiring a method for classifying a company as a "good" or "poor" investment choice for a given year. Although there is no definitive method for defining a market investment as "good" or "poor," in this study we use a method that is simple but yet objective, namely, if a company's stock prices ratio of two consecutive periods rose above the predefined market price ratio, it is classified as a "good" investment option; otherwise, it is classified as a "poor" investment option. Here the dependent variable denoted by y, where the label *1* is assigned to y if the prices ratio of two consecutive periods is greater than or equal to *1.50* and *-1* if it is lower than *1.50*. Hence the value $y = 1$ represents 'good' stock and $y = -1$ represents 'poor' stock. The stock analyst can assign the y values according to the predefined ratio which meets the targeted expected return.

Table 1. Estimation result based on the training data on Balance SVM model

Model	Objective value	Training classification accuracy(%)	Testing classification Accuracy(%)	Selected parameters (coefficients, constant)
Balance SVM	-40.56581 nSV=122, nBSV= 30	95.1613	91.5493	(-290.520 0.5406 -8.7218 46.0478 50.5916 -31.8660, -0.7106)
Imbalance SVM	-43.25200 nSV=123, nBSV=0 Total nSV= 123	100	95.7746	(468.635 0.8235 -16.9536 61.5862 73.5812 -81.6240, -0.5447)
Balance SVM	-40.2605 nSV=119, nBSV= 28 Total nSV=119	97.479	94.3662	(-320.181 0.3327 -3.5436 6.2903 -3.8666 -36.4044 7.4242 -2.0088, -0.6895)
Balance SVM	-57.8599 nSV = 90, nBSV = 42 Total nSV=90	79.8319	78.2609	(-1.1234 -0.1906 0.3095 -0.1430 0.0189 -0.3737, -0.7684)
Imbalance SVM	-57.8599 nSV = 90, nBSV= 42, Total nSV=90	79.8319	78.2609	(-1.1234 -0.1906 0.3095 -0.1430 0.0189 -0.3737, -0.7684)

The best parameter and hyper parameter model selection obtained through the model selection process which used *10*-fold cross validation method. The grid search procedure was integrated in the cross validation procedure in order to avoid over-fitting model Cawley & Talbot (2010). The model parameters searching was done in the given parameter intervals, namely [2^{-5}, 2^{5}] for the Radial Basis Function kernel parameter and [1, 10] interval for penalty cost parameter (C) searching. Figure 2 depicts the flow process of SVM model selection for imbalance data where figure (a) shows the proposed method and (b) the basic imbalance SVM model. The SVM classification model selections are carried out by running MATLAB script on the LIBSVM interface for several models. The five SVM models were including balance and imbalance SVM model implemented on the original data, imbalance SVM and imbalance models applied on the normalized features data and the balance SVM model implemented on the dummy addition feature variables. The usual measures of prediction model performance such as best fit model measurement (Accuracy) is used for model evaluation. The summary of the model evaluations interms of: Accuracy, value of objective function, number of Support vector (nSV), obtained coefficients and constant for the decision function is given in Table 1.

4. EMPIRICAL RESULT AND ANALYSIS

The weight assignments were according to Equation (5) to give the modified objective function for SVM optimization problem. As a result, the constrained Lagrange primal optimization formulation can be shown to yield two constraints as follows:

$$0 \le \alpha_i \le \mathrm{C}_+ \quad \text{if } \mathrm{y} = +1, 0 \le \alpha_i \le \mathrm{C}_- \text{ if } \mathrm{y} = -1 \tag{10}$$

Lagrangian coefficients α_i come from the minority class are much bigger than does come from the majority class. Hence the SVM algorithm can be expected to locate for the hyper plane to move closer to the minority class. The proposed dummy variable additions prior to the training process, can be expected to yield somehow similar effect. This is explained as follows. Assumed there are p feature variables and two dummy variables, where the first dummy represents the minority class, say the positive outcome and the second dummy represents the majority class, i.e., the negative class. The vectors data $\boldsymbol{x} = (x_1, x_2, \ldots, x_p, x_{d1}, x_{d2})$, containing *p* feature data and two dummy variables, where

$$x_{d1} = \begin{cases} 0 & if \ \ y_{i=-1} \\ \frac{n-}{n+} & if \ \ y_{i=+1} \end{cases}, \ x_{d2} = \begin{cases} 1 & if \ \ y_{i=-1} \\ 0 & if \ \ y_{i=+1} \end{cases}. \tag{11}$$

By assuming all others *p* variables to be zeros while the two remaining dummy variable values conditional to the class type, then it is possible to allocate the hyper plane to close to the ideal one, as the two class move away within the first quadrant, as shown in Figure 3. It is important to notice that weight assignment value for both dummy variables should not make a difficulty in the resulting coefficient estimates. Since they are dummies then their existences should not diminish the important contribution of other indicator variables. In order to achieve this expectation, then it is suggested to use lower value for $x_{d1}<1$ while maintaining the proportion (x_{d1}/x_{d2}) to be the same fraction as of n^-/n^+. It is worth to note that the resulting prediction model with the dummy addition variables yield the coefficients of optimistic decision and pessimist decision of the stock performance. These are given by the coefficient estimates of the first and second dummy variables. If the decision is optimistic then the value in the first dummy variable is equal to the weight value set up in

the first dummy variable while the other is set to 0. If it represents pessimist decision then the first dummy value is set to *0* and the second dummy value is set to the weight value.

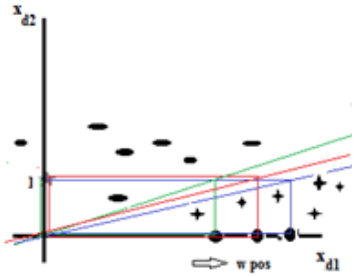

Figure 3. Separating hyper plane obtained by additions of two dummy variables

LIBSVM software Chang & Lin (2013) accommodates for imbalance data using two model parameterizations namely the 'w_1' and 'w_{-1}' flags (Hur &Weston, 2004). But there are some notes about how to choose appropriate weight value: 1) How to apply the weight values: Some practitioners suggest using small values less than 1 according to the proportion of number of data in the majority class and in the minority class hence there is an issue of weight magnitude scaling.

2) Based on our experiment; having all the set up parameter being supplied by the user, there is no different accuracy results for interchanging the weight assignments in the above weight flags.

3) The cross validation procedure to obtain the optimal penalty cost parameter through the grid search implementation is not efficient since after having obtained them, then they are multiplied again with the weights set in the 'w_1' and 'w_{-1}' flags.

By using results presented on Table 1, the model evaluation for determining the best model was obtained by considering three points: The objective function's value of the dual optimization problem should be as maximal as possible; the accuracy on the training data and testing data should be as maximal as possible. Based on these criterions, the obtained scores were as follows: Balance SVM model with dummy variables to have score = 3, Imbalance SVM model with score = 2, those all to have same scores =*1* were including Balance SVM which applied to normalized data set, imbalance SVM model applied on normalized data set and Balance SVM model applied on the original data set.

Prediction equation for balance model with dummy variable based upon known independent variables is obtained according to Equation (9) depending under the optimist or pessimist decision. For other prediction models, prediction equation (3) should be used.

Observation for the two models highest score models are as follows: For imbalance model, it can be seen that the negative coefficients belong to variables: ROE, EPS Ratio and NPM while positive coefficients belong to variables: DAR, PE ratio and PEG. For dummy balance model, it can be seen that the negative coefficients belong to variables: ROE, EPS Ratio, PEG and NPM while positive coefficients belong to variables: DAR and PE ratio. In this case, if DAR and PE ratio are high to some certain levels and ROE, EPS Ratio and NPM are low then it is likely to assign the stock to have 'good' performance. This result may be interpreted that the investors in the sharia stock market tend to value the debt ratio to have the positive contribution to the stock performances as well as the high market price paid for the earning yielded for each unit stock.

5. CONCLUSION

The proposed method to deal with imbalance data is better than imbalance SVM model in terms maximal objective function while for the two other measures, namely Accuracy in the training data and testing data are competitive hence this can be expected to help the decision process. Moreover, for the proposed technique of imbalance data prediction, the analyst can set the targeted price improvement that meets the expectation of return on stock investment through the determination of the price ratio of two consecutive periods to consider as 'good' or 'poor' stocks. It may be observed that six financial ratios namely: ROE, DAR, EPS rate of growth, PER, PEG ratio and NPM can classify companies up to 95.7746% level of accuracy into two categories ("good" or "poor") based on the predefined growth price ratio, which in this case was 1.50. Hence the proposed method can be considered as a promising tool for collecting the 'good' stocks in a better precision.

In this study, the annual data were taken into consideration hence the last known annual observations of stock prices were compared with those of the previous year to determine the performance. In further studies, data for each three-month period can be used, and different criteria can be defined, for evaluating stock performance. This study used financial ratios as the only factor affecting share prices, but there may be various other economic and management factors that may also influence share's performances.

Acknowledgement

Thanks to the Ministry of Higher Education of the Republic of Indonesia for the research funding under the grant number: 157/SP2H/PL/DT/LITABMAS/V/2013.

REFERENCES

1. Harvey, C.R. **Predictable risk and returns in emerging markets**, The Review of Financial Studies *8*, 773–816, 1995.

2. Kato, K., W. Ziemba and S. Schwartz. **Day of the week effects in Japanese stocks**, In: E. Elton and M. Grubber, eds., Japanese Capital Markets, New York: Harper & Row, 1990.

3. Jung, C., and R. Boyd, **Forecasting UK stock prices**, Applied Financial Economics 6, 279–86, 1996.

4. Lee, S., **Application of likelihood ratio and logistic regression models to landslide susceptibility mapping using GIS**. Environment Management 34(2), 223-232, 2004.

5. Zavgren, C. **Assessing the vulnerability to failure of American industrial firms: A logistic analysis**, Journal of Business Finance and Accounting 12(1): 19-45, 1985.

6. Cheng, W, L. Wanger, and Ch. Lin. **Forecasting the 30-year US treasury bond with a system of neural networks**, Journal of Computational Intellegence in Finance 4, 10-6, 1996.

7. Ogut, Hulisi, et al, **Detecting stock-price manipulation in an emerging market: The case of Turkey**, Expert System with Applications 36(9), 11944-11949, 2009

8. Van, E., and J. Robert. **The application of neural networks in the forecasting of share prices**. Haymarket, VA, USA: Finance & Technology Publishing, 1997.

9. Min J. H. and Y.C. Lee, **Busines Failure Prediction with Support Vector Machines and Neural Networks: A Comparative study**, Korea, 2004.

10. Min, Jae H., and Chulwoo Jeong. **A binary classification method for bankruptcy prediction**, Expert Systems with Applications 36(3), 5256-5263, 2009.

11. Kumar P.R., and V. Ravi**, Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review**, European Journal of Operation Research 180: 1-28, 2007.

12. Mostafa, Mohamed M**., Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait**, Expert Systems with Application 37(9). 6302- 6309, 2010.

13. Guresen, Erkam, et al. **Using artificial neural network models in stock market index prediction**, Expert System with Application 38 (8), 10389-1039, 2011.

14. Li, Hui, et al., **Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods**, Expert System with Application 37(8), 5895-5904, 2010.

15. Swiderski, Bartosz, et al. **Multi-stage classification by using logistic regression and neural networks for assessment of financial condition of company**, Decision Support System 52(2), 539-547, 2012.

16. Hsu C.W., Chang C.C., and Lin C.J., **A Practical Guide to Support Vector Classification**, 2009, http://www.csie.ntu.edu.tw/~cjlin., March 2013.

17. Hofmann T., Scholkopf B. and Smola A. J., **Kernel Methods In Machine Learning**, Institute of Mathematical Statistics, 2008.

18. C. Cortes and V. Vapnik. **Support-vector network**. Machine Learning, 20:273-297, 1995.

19. Anand A., Pugalenthi G., Fogel G.B, and Suganthan P.N., **Amino Acids: An Approach for classification of highly imbalanced data using weighting and undersampling**, Vol. 39, hal. 1385-1391, Springer Verlag, 2010.

20. Akbani R., Kwek S., Japkowicz N. **Applying support vector machine to imbalanced datasets**, In Proc 15th European Conference on Machine Learning (ECML) 2004.

21. Hur A.B.,Weston J., **A User's Guide to Support Vector Machines**, Departement of Computer Science Colorado State University, USA, 2008.

22. Han H., Wen-Yuan Wang, Bing-Huan Mao, **Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets,** Learningll ICIC 2005, Part I, LNCS 3644, pp. 878-887, 2005.

23. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. **SMOTE: Synthetic minority over-sampling techniquel**. Journal of Artificial Intelligence Research 16, 321-357, 2002.

24. Ganganwar, V., **An overview of classification algorithms for imbalanced datasets** in International Journal of Emerging Technology and Advanced Engineering:, Vol. 2, No. 4, April, ISSN 2250-2459, 2012.

Appendix 1

Sample Data Set (142 Observations)

Code	ROE	DAR	EPS 2010/ EPS 2009	P/E = P/ EPS 2010	PEG	NPM	Dplus	Dmin	sign
ASII	42.65	0.48	1.43	15.38	0.01	11.05	0	0.07	-1
TLKM	48.21	0.43	1.02	13.9	1.37	16.81	0	0.07	-1
PGAS	58.14	0.53	1	17.18	40.91	31.57	0	0.07	-1
SMGR	39.33	0.22	1.09	15.43	0.3	25.33	0	0.07	-1
ADRO	27.18	0.54	0.51	36.95	-0.55	8.94	0	0.07	-1
INDF	32.37	0.47	1.42	14.49	0.15	7.69	0	0.07	-1
BMTR	14.95	0.52	3.09	18.17	1.65	37.96	0.25	0	1
INTP	32.49	0.15	1.17	18.21	0.14	28.95	0	0.07	-1
ASRI	14.95	0.52	3.09	18.17	1.65	37.96	0.25	0	1
ITMG	38.47	0.34	0.58	31.29	-0.03	12.24	0	0.07	-1
MNCN	21.5	0.34	1.89	17.76	0.71	15.04	0	0.07	-1
KLBF	32.95	0.18	1.38	25.65	0.73	12.58	0.25	0	1
EXCL	33.02	0.57	1.69	15.6	0.11	16.56	0	0.07	-1
JSMR	19.07	0.56	1.2	19.52	0.66	27.26	0	0.07	-1
AKRA	17.6	0.63	0.94	21.11	-3.79	2.55	0.25	0	1

UNVR	112.19	0.53	1.11	37.17	0.83	17.2	0	0.07	-1
PTBA	40.83	0.26	0.74	26.34	-0.08	25.4	0	0.07	-1
BSDE	10.8	0.37	0.8	39.96	-7.02	15.92	0	0.07	-1
LPKR	9.33	0.49	1.08	28.03	15.07	16.81	0	0.07	-1
LSIP	30.34	0.18	1.46	16.97	0.07	28.76	0	0.07	-1
CPIN	63.21	0.31	2.75	11.5	0.01	14.66	0	0.07	-1
BKSL	2.52	0.14	9.16	47.75	23.41	14.76	0.25	0	1
INCO	34.61	0.23	2.44	12.33	0.05	34.27	0	0.07	-1
MAPI	18.77	0.6	1.23	22.07	0.99	14.77	0.25	0	1
ANTM	23.72	0.22	2.79	13.89	0.12	19.25	0	0.07	-1
INDY	7.93	0.52	1.06	31.81	3.52	20.52	0	0.07	-1
CMNP	21.52	0.37	4.32	9.13	0.08	39.75	0	0.07	-1
AALI	41.1	0.15	1.21	20.46	90.54	22.8	0	0.07	-1
SMCB	16.83	0.35	0.92	18.2	-2.07	13.9	0	0.07	-1
WIKA	26.27	0.7	1.47	14.23	0.94	4.73	0	0.07	-1
TSPC	24.17	0.28	1.36	15.71	0.55	9.52	0	0.07	-1
SCMA	48.54	0.41	1.85	12.94	0.1	27.5	0.25	0	1
INTA	28.54	0.73	2.22	290.03	2.75	4.53	0	0.07	-1
TINS	26.82	0.29	3.02	14.59	0.12	11.37	0	0.07	-1
GJTL	31.77	0.66	0.92	9.64	-0.45	8.43	0	0.07	-1
SGRO	29.57	0.25	1.6	13.26	0.15	19.54	0	0.07	-1
MYOR	33.06	0.54	1.3	17.03	0.12	6.7	0	0.07	-1
HEXA	34.42	0.49	1.36	22.19	0.26	8.39	0	0.07	-1
MPPA	0.79	0.37	16.37	1.43	0	67.89	0	0.07	-1
ASGR	33.94	0.52	1.77	7.85	0.21	7.56	0.25	0	1
BYAN	36.73	0.64	5.44	80.98	0.45	8.47	0	0.07	-1
DVLA	24.02	0.25	0.77	11.84	-0.39	11.93	0	0.07	-1
SCBD	17.41	0.27	0.29	13.63	-0.24	6.88	0.25	0	1
TRAM	12.27	0.42	1.06	51.18	76.38	26.09	0.25	0	1
BWPT	29.46	0.57	1.45	21.36	1.13	34.2	0	0.07	-1
JPFA	46.74	0.5	1.18	6.79	0.1	6.87	0	0.07	-1
TLKM	34.2	0.41	1.34	9.18	0.05	21.53	0	0.07	-1
UNTR	28.3	0.41	1.35	16.75	0.04	10.65	0	0.07	-1
PGAS	44.54	0.45	0.99	12.49	-3.99	31.5	0.25	0	1
SMGR	34.83	0.26	1.09	17.16	0.31	24.18	0.25	0	1
ADRO	41.05	0.57	2.3	11.17	0.12	14.03	0	0.07	-1
INDF	20.1	0.41	1.7	8.06	0.03	11.07	0	0.07	-1
BMTR	4.49	0.36	0.34	70.21	-2.51	12.37	0.25	0.07	1
INTP	29.92	0.13	1.12	20.05	0.2	25.93	0	0.07	-1
ASRI	24.08	0.54	2.08	13.64	0.78	43.64	0	0.07	-1
ITMG	67.54	0.32	2.73	10.21	0	23.15	0	0.07	-1
MNCN	22.1	0.22	0.54	30.76	-1.27	21.4	0.25	0	1

KLBF	30.5	0.21	1.2	33.13	1.33	126.66	0	0.07	-1
EXCL	28.22	0.56	0.98	13.64	-1.8	15.12	0	0.07	-1
JSMR	18.68	0.57	1.11	21.6	1.15	26.64	0	0.07	-1
HRUM	65.18	0.23	2.07	10.82	0.03	23.43	0	0.07	-1
AKRA	20.7	0.57	1.63	22.59	0.44	2.72	0	0.07	-1
UNVR	151.45	0.65	1.23	34.45	0.34	17.74	0	0.07	-1
ENRG	10.8	0.37	2.57	2.94	0.08	36.07	0	0.07	-1
PTBA	49.71	0.29	1.54	12.97	0.03	0.29	0	0.07	-1
BSDE	10.8	0.37	2.57	15.57	0.44	36.07	0	0.07	-1
LPKR	10.47	0.48	1.03	26.28	31.29	13.84	0.25	0	1
LSIP	35.8	0.14	0.4	7.35	-0.02	44.61	0	0.07	-1
CPIN	48.06	0.3	0.11	14.93	-0.01	13.16	0.25	0	1
ICBP	25.63	0.3	1.21	14.68	0.24	10.66	0.25	0	1
BKSL	3.41	0.13	1.9	60.89	29.56	29.82	0	0.07	-1
INCO	25.57	0.27	0.77	10.5	-0.12	26.86	0	0.07	-1
MAPI	27.01	0.59	1.78	23.84	0.25	6.08	0	0.07	-1
ANTM	23.85	0.29	1.14	8	0.32	18.6	0	0.07	-1
INDY	7.35	0.58	1.55	9.49	0.12	22.99	0	0.07	-1
CMNP	17.89	0.32	1.18	9.55	0.35	43.93	0	0.07	-1
AALI	39.55	0.17	1.24	11.72	38.31	15.61	0	0.07	-1
SMCB	20.37	0.31	0.76	19.91	-0.47	23.19	0	0.07	-1
WIKA	28.37	0.73	2.06	9.17	0.27	14.02	0.25	0	1
TSPC	24.3	0.28	1.2	19.62	0.92	5.19	0	0.07	-1
SCMA	80	0.4	0.58	28.44	-0.15	10.13	0	0.07	-1
INTA	31.46	0.86	0.29	10.58	-0.08	39.56	0	0.07	-1
TINS	27.58	0.3	0.95	8.57	-0.85	4.01	0	0.07	-1
GJTL	19.31	0.62	1.14	11.05	0.33	10.25	0	0.07	-1
SGRO	29.7	0.27	1.22	10.23	0.2	7.99	0	0.07	-1
MYOR	25.84	0.63	1	22.6	68.49	17.49	0	0.07	-1
HEXA	39.61	0.52	1.62	17.17	0.09	5.12	0	0.07	-1
MPPA	2.89	0.45	0.02	42.52	-0.04	10.52	0	0.07	-1
TURI	29.26	0.42	1.2	10.35	1.06	1.35	0.25	0	1
ASGR	32.7	0.51	1.18	11.01	0.71	3.9	0	0.07	-1
DVLA	22.85	0.22	1.09	10.66	1.19	15.78	0.25	0	1
SCBD	4.9	0.25	0.95	37.24	33.25	12.44	0	0.07	-1
TRAM	10.15	0.71	1.23	74.9	27.14	10.56	0	0.07	-1
BWPT	30.15	0.6	1.31	11.18	0.59	26.11	0.25	0	1
JPFA	23.04	0.54	0.68	12.17	-0.08	36.07	0.25	0	1
TLKM	36.17	0.4	1.19	10.6	0.07	23.84	0	0.07	-1
UNTR	23.05	0.36	1	13.34	15.16	23.84	0	0.07	-1
PGAS	48.76	0.4	1.43	14.39	0.13	10.47	0	0.07	-1
SMGR	34.61	0.32	1.24	23.5	0.14	35.49	0	0.07	-1

INDF	15.03	0.42	0.77	13.35	-0.1	12.49	0	0.07	-1
BMTR	13.56	0.3	6.85	24.87	0.3	10.32	0	0.07	-1
INTP	32.13	0.15	1.32	18.71	0.06	21.5	0	0.07	-1
ASRI	28.41	0.57	1.83	10.97	0.39	27.55	0	0.07	-1
ITMG	58.97	0.33	0.83	12.24	-0.02	49.71	0	0.07	-1
MNCN	24.13	0.21	3.08	28.14	0.47	17.49	0	0.07	-1
KLBF	24.12	0.21	0.82	42.47	-1.58	27.78	0	0.07	-1
EXCL	24.41	0.57	0.97	18.37	-1.76	13.07	0	0.07	-1
JSMR	17.22	0.59	0.95	29.36	-3.31	13.09	0	0.07	-1
HRUM	47.54	0.26	0.77	12.03	-0.08	22.55	0	0.07	-1
AKRA	14.04	0.57	1.15	27.04	1.38	16.86	0	0.07	-1
UNVR	96.01	0.59	0.88	43.54	-0.65	0.57	0	0.07	-1
ENRG	-0.62	0.5	0.21	63.45	12.82	17.96	0.25	0	1
PTBA	45.99	0.33	0.74	15.41	-0.04	2.26	0	0.07	-1
BSDE	10.61	0.37	0.9	21.38	-3.56	34.48	0.25	0	1
LPKR	9.82	0.51	3.19	12.47	0.23	34.48	0.25	0	1
LSIP	21.85	0.17	0.54	14.38	-0.1	48.4	0	0.07	-1
CPIN	39.52	0.32	1.04	29.25	4.63	26.65	0	0.07	-1
ICBP	20.29	0.31	0.84	26.19	-0.47	15.5	0.25	0	1
BKSL	4.32	0.19	1.33	32.69	23.02	10.71	0.25	0	1
INCO	2.28	0.26	0.09	84.05	-0.3	33.2	0	0.07	-1
MAPI	19.45	0.63	0.8	38.51	-0.9	4.17	0	0.07	-1
ANTM	7.99	0.38	0.33	19.5	-0.14	5.27	0	0.07	-1
INDY	4.63	0.56	0.08	7.6	-0.04	8.81	0	0.07	-1
CMNP	15.77	0.31	0.85	11.13	-0.43	18.79	0.25	0	1
AALI	37.64	0.25	0.98	13.78	-0.48	6.57	0	0.07	-1
SMCB	1.97	0.59	0.45	33.46	15.42	21.22	0	0.07	-1
WIKA	19.3	0.76	1.61	28.82	1.47	0.55	0.25	0	1
TSPC	20.71	0.28	0.96	29.83	-5.68	4.92	0	0.07	-1
SCMA	54.81	0.24	0.34	27.3	-0.15	11.61	0	0.07	-1
TINS	12.43	0.3	5.52	1.56	0	1.53	0	0.07	-1
GJTL	18.92	0.59	0.72	11.94	-0.16	6.23	0	0.07	-1
SGRO	13.04	0.34	0.42	24.84	-0.15	7.25	0	0.07	-1
MYOR	23.48	0.63	1.07	29.7	0.69	10.84	0.25	0	1
HEXA	26.71	0.61	0.32	18.51	-0.05	6.72	0.25	0	1
MPPA	2.66	0.45	1.41	37.64	4.23	10.63	0.25	0	1
TURI	30.87	0.47	1.21	13.52	1.13	2.13	0	0.07	-1
ASGR	35.84	0.49	1.18	12.3	0.67	3.91	0	0.07	-1

[1] Retno Maharesi and Sri Hermawati are both lecturers at Information Technology and Management Department of Gunadarma University, Jakarta, Indonesia. The highest academic degree they own are PhD in Computer Science and Economic respectively.