Aktüerya Derneği

İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 14, 2021, 1, 14-29

Geliş/Received:26.01.2021, Kabul/Accepted: 03.5.2021

Araştırma Makalesi / Research Article

Applied integration of time series and multi-variable regression algorithms

Fatih Koyuncu

Department of Mathematics,
Ankara Yıldırım Beyazıt University, 06010,
Ankara, TURKEY
fatih@ybu.edu.tr
orcid.org/0000-0001-6351-4787

Ahmet Yücel

Department of Banking and Finance,
Ankara Yıldırım Beyazıt University, 06950,
Ankara, TURKEY
ayucel@ybu.edu.tr
orcid.org/0000-0002-2364-9449

Abstract

Time Series (TS) based prediction models generate prediction based data that is supposed to be similar to the future data at a certain level. In this study, we designed new modeling that increases the prediction performance of the TS algorithm. The main purpose of the new modeling is to integrate the Multivariate-Adaptive-Regression-Splines (MARSplines) algorithm into the TS algorithm. Five-year Tokyo Stock Exchange data is analyzed as a case study to apply the relevant models. The results show that the new regression-based approach significantly improves the prediction performance of the time series algorithm.

Keywords: Exponential correction, MARS, multivariate adaptive regression, stock price, time series analysis.

Öz

Zaman Serisi (ZS) tabanlı tahmin modelleri, belirli bir düzeyde geçmiş verilere benzer fonksiyonel dağılıma sahip, tahmine dayalı veriler üretir. Bu çalışmada, ZS algoritmasının tahmin performansını artıran yeni bir modelleme tasarlanmıştır. Yeni modellemenin temel amacı, Çok Değişkenli Uyarlamalı Regresyon Katmanları (MARSplines) algoritmasını ZS algoritmasına entegre etmektir. Beş yıllık Tokyo Menkul Kıymetler Borsası verileri, ilgili modelleri uygulamak için bir vaka çalışması olarak analiz edilmiştir. Sonuçlar, yeni regresyon temelli yaklaşımın ZS algoritmasının tahmin performansını önemli ölçüde geliştirdiğini göstermiştir.

Anahtar Kelimeler: *Üstel düzeltme, MARS, çok değişkenli uyarlamalı regresyon, hisse senedi fiyatı, zaman serisi analizi.*

1. Introduction

Many countries have stock exchange markets (SEMs) to make investments and to observe sector-based value movements. Mostly SEM is an open-source platform that shares information in synchronization with the whole world. The opening/closing hours of the SEM vary according to the local times. In SEM trading within specified local hours, up/down value movements are observed hundreds of times during the trading period. The value at which the stock exchange starts the day is called the "opening value", and the value it completes the day is called the "end of day value". By comparing the end-of-day values of consecutive days, the value movements of the stock exchange in daily, weekly, monthly, or yearly periods can be observed.

During the day, market values can change at any time depending on instant purchases/sales and political/social developments. For this reason, many irrational value movements directly related to social developments can be found in data sets taken from stock exchanges. Skills such as detecting the trends in the markets, making the right timing in investment preferences, predicting the market's reaction in advance are very critical for investors who want to evaluate their investments on this platform. It is very difficult to create a model that can make statistically accurate predictions on this platform, which produces non-parametric data that is open to irrational effects. For this reason, there is a need for more advanced algorithms that support standard approaches with new methods.

The basic working principle of the TS prediction models is based on the assumption that existing trends or patterns in the historically older part of the data continue to exist in the current or future part of the data. In other words, existing data sets have information about future data at certain levels. The TS algorithms, which are widely applied in the field of data mining and have a very high level of prediction success, provide the ability to detect statistical connections present in historical data and to make predictions about future data using this information.

This study proposes a prediction algorithm that expresses the distribution according to time functionally. The emphasis of the experimental study is on a continuous scale review of market-generated data with a public license. The contributions to the literature and strengths of the study are as follows: (i) In recent years, there has been a huge increase in the number of analysis methods that are highly accurate but require a deeper understanding of machine learning methods and the external intervention of researchers. Therefore, the models based on those methods can only be created experimentally. The proposed approach gives accurate results at higher levels while reducing the external intervention of researchers and the amount of data required for training models. (ii) It makes it possible to easily analyze data sets from different fields with any size and number of variables, without the need for advanced knowledge of machine learning techniques by researchers.

The ability to make successful predictions about future data is critical, especially in the field of finance. The TS and similar machine learning-based statistical tools can predict future data at certain levels. In this study, 5-years data obtained from the Tokyo SEM is examined and after experimental studies on the relevant data, a new algorithm is developed to improve the current prediction potential.

MARSplines is a data mining technique that offers sufficient flexibility and precision for fast estimation of both continuous and binary categorical variables. Also, MARSplines models develop a functional relationship with regression data. The main advantage of the MARSplines model is that it causes fewer variables to interact with these models [9].

In the field of economy, the most important factor affecting the decision of investors in investment preferences is the possibility of realization of future projections of experts and general expectations. For this reason, predicting future data with high accuracy will determine direct investment preferences. For this purpose, many statistical forecasting models are presented to guide investors in the economy and finance sector. One of the most important of these models is predictive modeling based on time series. Al-Idrisi is applying a special exponential smoothing TS prediction algorithm to predict the share prices of companies that are publicly bought in the Saudi SEM. In the study, the success of the applied model with the regression-integrated hybrid method is compared [5].

In another study from the domain of economics and finance, Neto et. al offers a similar hybrid model that combines the TS and artificial neural network (ANN) algorithms to predict the stock prices of some companies in Brazil. The main idea of the algorithm is to integrate one of the TS models which has the highest correlation with the dependent (stock prices) variable into the ANN model to predict the dependent variable. In this study, several methods are compared in terms of their performances. It is observed that the best predictive performance is obtained from trained neural network models [6].

A new chart pattern recognition model that integrates MARSplines and RNN (recurrent neural network) methods (MARS-RNN) is introduced by Kao et. al. This model is an important tool in detecting existing patterns in the production process control charts and providing a valuable reference for developing new strategies with findings. The performance of the created model has been compared with the Random Forest (RF), ANN and MARSplies integrated models. It has been observed that the new model performs better than other models in both independent variable selection and prediction processes [11].

Another hybrid model that combines regression and time series algorithms is also presented by Guolo and Varin. In this study, many regression methods, including the MARSplines algorithm, were tested by integrating them into time series. The main motivation for the study is to observe the progress of epidemics and to predict future points. In this way, early preparation will be possible for possible future scenarios of outbreaks. In the study, the performances of the MARSplines based hybrid model and the Back Propagation Neural Network (BPNN) model were compared. Results show that the MARSplines hybrid model can make more accurate predictions than BPNN [1].

With the high performance of BPNN, it is a method that has the power to be sufficient in many studies alone. Therefore, the comparison of the hybrid model with BPNN was made for a purpose in this regard. In a similar study, Arasu compares the stock index prediction success of the BPNN and MARSplines methods. It has been concluded that when the MARSplines method is applied to time series, it has a higher accuracy success than BPNN [2].

Time series models are a highly successful method to predict the future behavior of events that occur in cyclic models of processes that repeat at periodic intervals. One of the most important advantages of the method is that the time series have ready models. However, the parabolic movements of nonlinear data and high-frequency events are ignored in this method [10].

One of the main reasons why the MARSplines algorithm is preferred is the high prediction success of its method in nonlinear models due to its adaptive and flexible model structure. Another reason for the preference is that it can be easily applied to large-sized data with numerous independent variables [3].

In another study conducted on non-linear data, the ratio of particulate matter (PM100) in the air throughout Ankara province in a period covering the years 1993-2017 is examined. Since it has a content that directly affects human respiratory quality and therefore human life, a very sensitive modeling process is needed, just like the studies in the field of medicine. The effects of seasonality and periodicity in the data are taken into consideration. For this reason, the designed model is based on the support of the linear TS integrated with harmonic regression models [15].

A similar weather study is conducted on snowfall modeling of a region in Iran. To model, the support vector machine (SVM) integrated with MARSplines, and the RF integrated with TS methods are used together. The results showed that the RF integrated with the TS model outperformed [9].

The MARSplines method consists of step-by-step, interconnected basic functions, and the number of basic functions and related parameters in each step (spline) is automatically determined by the data. In this method, the data are split into sections over and over for each step, and models based on continuous variables are created according to basic function derivatives within each section [3].

In the study of Lee et. al., by integrating support vector regression (SVR) and autoregressive moving average (ARMA) regression-based models, a hybrid model was created that can determine the change point for time series [13]. Especially in the medical field, when conducting an analysis based on TS, the detection of change points in the timeline is vital [12].

In the analysis models based on TS applied in the medical field, data content is directly related to human health and requires a very sensitive model formation process. Having a similar and repetitive pattern of vital processes positively affects the success of TS models. Traditional vital signs are controlled within predetermined value ranges. Warning alarm systems operate when the relevant values fall outside these ranges. Looking at the course of the process with TS, potential alarm situations in the future can be predicted. However, with physiological and medicated interventions in the process leading to the alarm state, the situation can be controlled before reaching the alarm level. We can express these stages as peak or turning notes. TS serves as an important tool for observing the effects of physiological and medicated interventions in the relevant process. In the study of Grillenzoni et. al., he introduced a new approach that integrates time series with adaptive regression methods, aiming to perceive the situations expressed as peak or turning point in advance. Thanks to this method, the change points of local trends in a series of graphs can be determined instantly [12].

Also, in the model produced by the MARSplines algorithm in the last step, the relationship pattern and level between the independent variables can be obtained. This situation highlights the MARSplines method among other methods as an easy method to understand and interpret, especially for multivariable complex models [2].

In other words, the MARSplines method is a method that provides tremendous convenience for nonlinear and complex large data sets, taken from many different fields from economics to biology. Leathwick, for example, studied a very complex dataset where fifteen different fish species examined their relationship to each other and the environment. It uses this method to generate a model that predicts the future behavior of the species [4].

Multivariate time series (MTS) integrated regression models are used in deep learning (DL) based data mining studies in many areas from finance to health. Mode et. al. applies DL and MTS integrated regression model to increase the prediction performance of TS models based on cybersecurity data [14].

Deep Neural Networks (DNNs) have increasingly widespread use in data mining with their high accuracy performance. In another study, a new approach is introduced that integrates the Generalized Linear Model (GLM), Seasonal Automatic Stress Integrated Moving Average model

(SARIMA), and Automatic Stress Integrated Moving Average (ARIMAX) methods with DNN using classical regression methods. The superior performance of this approach has been demonstrated when compared to existing hybrid models in the literature [16].

The TS prediction models aim to generate predictions about future data by analyzing the pattern created by past observations. This is important to support decision-making about future issues in many areas. In his study, Ilic, which is another hybrid approach based on time series and regression algorithms, developed a linear regression (EBLR) algorithm that explains the errors of the classical TS model through regression trees and provides accuracy-enhancing model learning in later steps. The model consists of two stages. In the first stage, estimates are made based on a basic time series model. In the second step, the errors of the current model are calculated with a regression tree [17].

Thanks to the methods and motivations provided by the studies mentioned in this section, a new hybrid approach has been developed in which MARSplines is integrated into the exponential smoothed TS algorithm. The methodology of the study and details of the applied methods are explained in Section-2. In Section-3, the application of the stated methods and methodology on data is given. In the last section, the main findings and results obtained from the research are expressed.

2. Methodology

In this study, time-series additive exponential smoothing estimation algorithm is integrated with the multivariate adaptive regression algorithm. Thus, with this hybrid regression model, we aim to increase the prediction ability of the classical time series algorithm.

2.1. *Additive Exponential Smoothing Time-Series Analysis*

A panel dataset usually has a recurring pattern over some time. The pattern that periodically repeats itself is called the seasonal effect of a time series model. Regardless of possible pattern repetitions, the specified time intervals are called specified periods depending on the date. The exponential smoothing algorithm gives lower weights to the specified periods when the length of the period increases [7]. The seasonal effect of a specified period assigned by the exponential smoothing model is calculated as follows.

$$y_t = \mu_t + \beta_t \mathrm{t} + S_{t,p} + a_t \tag{1}$$

where μ_t is the mean value of each period, β_t is the slope effect (trend) parameter, and $S_{t,p}$ is the seasonal effect of the p seasons (p =$1,2, \dots ,P$) in a year, and a_t ($t = 1,2, \dots ,T$) is the white noise error parameter. Expressed parameters are not fixed. Each parameter value may change over time.

If a smoothing model has no trend, then β_t = 0 for all t. Similarly, if a smoothing model has no seasonal effect, then $S_{t,p} = 0$ for all p. So, the smoothing parameter given above are calculated with the following equations.

$$L_t = \alpha(y_t - S_{t-P}) + (1-\alpha)(L_{t-1} + T_{t-1}) \tag{2}$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1} \tag{3}$$

$$S_t = \delta(y_t - L_t) + (1-\delta)S_{t-P} \tag{4}$$

where $\delta, \alpha,$ and γ are seasonal effect, level, and trend weights of the smoothing, respectively. Also let the smoothed level, smoothed slope (trend) and smoothed seasonal estimators be L_t, T_t, and S_t such that L_t estimates μ_t, T_t estimates β_t, and S_t estimates $S_{t,p}$, at time t. Let h be the number of estimated cases. Then the h-step-ahead esimation equation is

$$\hat{y}_{t+h} = L_t + hT_t + S_{t-P+h} \ , \quad h = 1, 2, \ldots \tag{5}$$

where $\hat{y}_{t+h}$ is the estimated h cases of y, hT_t is the estimated h additive slope, and S_{t-P+h} is the h estimated smoothed seasonal factor [8].

2.2. *Multivariate Adaptive Regression Splines (Marsplines)*

The MARSplines algorithm was produced by Friedman in 1991[3]. As computerized statistical data analysis tools have become widespread, MARSplines has become an increasingly popular method. The basic logic of the algorithm is to predict the values of a continuous dependent variable with the help of continuous or categorical independent variables. MARSplines algorithm does not produce a functional estimator based on the relationship between the dependent and independent variables as in classical regression models but instead offers a nonparametric regression model. The algorithm aims to describe the possibly existing relationship between variables by a method using basic functions in addition to the coefficients generated from classical regression data. The general procedure of the method consists of two steps; First, the data is split into two equally sized subsets. In the next step, a hybrid regression model is generated that includes the basic function for each subset. The MARSplines regression model is a combination of these hybrid models. Classical regression models generally offer more successful approaches for such datasets with a linear distribution. Thanks to this expressed feature of the MARSplines method, very successful models can be created even for datasets having non-linear distribution. It will be quite difficult to create similar models with parametric methods. The MARSplines algorithm uses double-sided functions similar to basic functions defined for linear or nonlinear distribution. The purpose of these functions is to create the alternative path from the dependent variable to the independent variables with the best possible approximation. Also, the functions generate estimations for the dependent variable. MARSplines model is given with Eq. (6) [3]:

$$y = f(X) = B_0 + \textstyle\sum_{i=1}^{S} B_i \, h_i(X) \tag{6}$$

where *S* is the number of splines in the model, *y* is prediction for the dependent variable, *X* is the index of the independent variables, B_0 is estimated intercept parameter, B_i is the estimated parameters and the $h_i(X)$ is the basis function.

Depending on the number of seasons and periods, basis functions are interchanged through a MARSplines model several times for finding the best match to the least-squares goodness-of-fit criterion. At the end of this process, the model automatically decides on the importance of the independent variables and their mutual interactions. As with all the nonparametric models, the MARSplines model is adaptive and therefore causes an overfitting problem. For overcoming this problem, the model has a pruning method decreasing the number of basis functions for decreasing the complexity of the model. This pruning process makes MARSplines a very strong method for predictor selection [3].

The MARSplines algorithm’s predictor selection process is as follows: The first step of the process starts with a model having only the constant basis function. In the second step, the basis function definition space is scanned for the most successful variables and knots, and then the selected variables and knots are added into the model for minimizing the prediction error (goodness of fit measure). In the third step, the second step is repeated until obtaining the pre-determined maximum measurement for the goodness of fit. After an exact number of iterations, in the final step, the pruning method is applied based on functions and so the functions contributing least to the goodness of fit (least squares) are removed from the model [3].

2.3. *Integration of Time-Series and MARSplines Algorithms*

Table 1. Mathematical process of the integrated hybrid algorithm

$$y = f(X) \longleftarrow X \longleftarrow y_t \longleftarrow \left[\begin{array}{lllll} \mu_t & \longleftarrow & L_t & \longleftarrow & \alpha \\ \beta_t & \longleftarrow & T_t & \longleftarrow & \gamma \\ S_{t,p} & \longleftarrow & S_t & \longleftarrow & \delta \\ a_t & & & & \end{array} \right.$$

The mathematical process of the integrated hybrid algorithm is given in Table 1. The hybrid algorithm consists of adding steps of predictors into the MARSplines model such that each predictor is recalculated based the exponential smoothing time-series model. For this purpose, level (α), trend (γ) and seasonal effects (δ) of smoothing weight are determined. Then, according to these parameters, smoothed level (L_t), smoothed slope (T_t) (trend) and smoothed seasonal estimators (S_t) are calculated in such a way that L_t estimates μ_t, T_t estimates β_t, and S_t estimates $S_{t,p}$ at time t. As a result, the seasonal exponential smoothing model Eq. (1) is generated with these obtainded parameters. This process is repeated for each iteration. Then, since each y_t value will be obtained after the calculation based on the time-series algorithm, the equation Eq. (7) is obtained by replacing the X predictor parameter in the MARSplines model Eq. (6) [3].

$$y = f(y_t) = B_0 + \sum_{t=1}^{T} \sum_{i=1}^{S} B_i\, h_i(y_t) \tag{7}$$

As a result, the mathematical expression of the hybrid algorithm Eq. (8) is obtained as follows [3].

$$y = f(y_t) = B_0 + \sum_{t=1}^{T} \sum_{i=1}^{S} B_i\, h_i(\mu_t + \beta_t \mathrm{t} + S_{t,p} + a_t) \tag{8}$$

3. Experimental Work

The dataset of the study has been taken from Kaggle.Com, an open-access database, and statistical consulting platform. The original name of the data is ’Uniqlo (Fast Retailing) Stock Price‘. The daily opening value (Open), closing value (Close), daily highest value (High), daily lowest value (Low), daily trading volume (Volume), daily stock trading total value (Stock Trading) of the Tokyo Stock Exchange. The Stock_Trading variable contains the daily total purchase / sale values of the shares of the company Uniqlo, which are traded on the stock exchange. In addition, Date variable is the only categorical variable. The first case of the data is received on January 4, 2012, and the last case is received on January 13, 2017. There are only 7-days-data available from 2017. There are a total of 1233 cases in the dataset. The first 1226 of the cases are used as a training set and the remaining 7 cases are reserved as a testing set. The data is licensed under ‘Public Domain Dedication’ (CC0 1.0 Universal (CC0 1.0)).

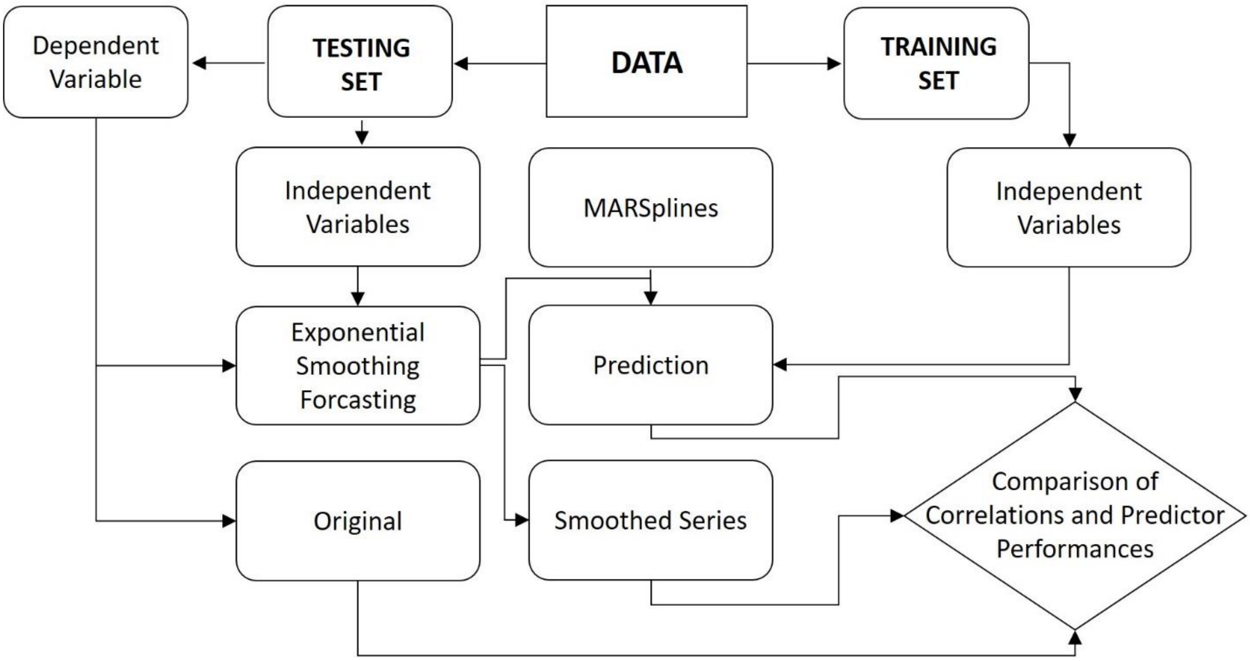

Figure 1. Methodology

To better observe the performance of the model, data that is open to irrational effects was preferred instead of data in which the values of consecutive days are linearly arithmetically similar to each other. For this purpose, stock market data was considered appropriate. The selection of the Tokyo stock exchange is random and does not have a specific purpose.

In this study, a new algorithm is designed by integrating a time series based prediction algorithm with a multivariate adaptive regression model. To observe the performance of the obtained algorithm, a situation analysis was made on the dataset of the Tokyo stock market. The dependent variable of the time series prediction model built on the integrated algorithm is determined as Stock_Trading. Except for the date variable, all other variables are determined as independent variables (predictors). The predictive success of the algorithm is evaluated according to its prediction proximity rates to the original values of the predicted cases on the testing set. For this purpose, the mean absolute residuals between the predictive and original values of the dependent variable are calculated. It can be concluded that the success of the model increases as the mean residuals approach zero.

The first estimation model is called as the Additive Exponential Smoothing Time-Series Analysis algorithm. Stock_Trading is identified as lock variable. Also the specified parameters; the additive season number (*t*) is 12 (corresponding to 12 months), the level smoothing weight is $\alpha = 0{,}01$, the seasonal smoothing weight is $\delta = 0{,}01$, and the trend smoothing weight is $\gamma = 0{,}01$.

The second prediction model is based on the new integrated algorithm. The modeling process is performed in three stages. In the first stage, the same independent variables (*Open, Close, High, Low, and Volume*) that are used in the Additive Exponential Smoothing Time-Series Analysis model, have been determined as Lock parameter one by one. In other words, the algorithm that is used to estimate the dependent variable in the first model is also applied to the independent variables in this step, so that the case values of the independent variables in the testing set are estimated. In the second stage, the MARSplines model is generated on the training set by using the predefined variables. The mathematical expression of the model is used as a reference function in calculating predictive values in the next stage. In the third stage, the predictive values obtained in the first stage are inserted in the function obtained in the second stage, and thus each case of the dependent variable is recalculated. Then, the absolute correlations between the predicted and original set are calculated.

In the final step, two separate multivariate regression models are built with the original and predicted sets, as stated above, to evaluate the performance of the algorithms.

3.1. *Results*

An exponential smoothing time series prediction model having *Stock_Trading* as a dependent variable is constructed by using the control (training) set. The parameter values that are applied when creating the model are the same as the initial constants. In this way, all of the data in the testing set is recalculated according to this model. The values calculated based on the model are given in Table 3 ($Notation\text{: } E^{+}N = 10^{N}$ and $E^{-}N = 10^{-N}$). The graph of the model is given in Fig. 2.

Table 2. Time-Series prediction

TEST	OPEN	HIGH	LOW	CLOSE	VOLUME
OBSERVED	4,25E^{+}4	4,33E^{+}4	4,25E^{+}4	4,33E^{+}4	6,48E^{+}5
	4,33E^{+}4	4,33E^{+}4	4,25E^{+}4	4,26E^{+}4	5,17E^{+}5
	4,05E^{+}4	4,10E^{+}4	3,97E^{+}4	3,97E^{+}4	1,44E^{+}6
	3,86E^{+}4	3,89E^{+}4	3,82E^{+}4	3,87E^{+}4	1,20E^{+}6
	3,87E^{+}4	3,89E^{+}4	3,85E^{+}4	3,86E^{+}4	5,46E^{+}5
	3,83E^{+}4	3,85E^{+}4	3,79E^{+}4	3,80E^{+}4	8,01E^{+}5
	3,89E^{+}4	3,94E^{+}4	3,82E^{+}4	3,84E^{+}4	1,32E^{+}6
PREDICTED (TIME SERIES)	4,31E^{+}4	4,33E^{+}4	4,25E^{+}4	4,27E^{+}4	5,07E^{+}5
	4,29E^{+}4	4,32E^{+}4	4,25E^{+}4	4,27E^{+}4	5,47E^{+}5
	4,29E^{+}4	4,31E^{+}4	4,26E^{+}4	4,27E^{+}4	4,58E^{+}5
	4,28E^{+}4	4,31E^{+}4	4,24E^{+}4	4,26E^{+}4	4,31E^{+}5
	4,26E^{+}4	4,29E^{+}4	4,22E^{+}4	4,25E^{+}4	4,87E^{+}5

	4,27E^{+}4	4,31E^{+}4	4,25E^{+}4	4,27E^{+}4	4,55E^{+}5
	4,27E^{+}4	4,30E^{+}4	4,24E^{+}4	4,25E^{+}4	4,22E^{+}5
RESIDUALS	-6,04E^{+}2	1,53E^{+}0	-4,77E^{+}1	5,91E^{+}2	1,41E^{+}5
	3,22E^{+}2	1,46E^{+}2	1,12E^{+}0	-1,40E^{+}2	-3,07E^{+}4
	-2,41E^{+}3	-2,10E^{+}3	-2,86E^{+}3	-3,00E^{+}3	9,78E^{+}5
	-4,16E^{+}3	-4,21E^{+}3	-4,28E^{+}3	-3,92E^{+}3	7,66E^{+}5
	-3,88E^{+}3	-4,05E^{+}3	-3,69E^{+}3	-3,96E^{+}3	5,89E^{+}4
	-4,42E^{+}3	-4,67E^{+}3	-4,53E^{+}3	-4,73E^{+}3	3,46E^{+}5
	-3,81E^{+}3	-3,59E^{+}3	-4,12E^{+}3	-4,07E^{+}3	9,00E^{+}5

A new time series model is composed for the recalculated parameters (*Open*, *Closed*, *High*, *Low*, *and Volume*), and predicting is performed for the dependent variable. Prediction values calculated based on the time series algorithm are given in Table 2. Also, the observed values existing in the testing set, the prediction values, and the differences (residuals) between these two values are given in Table 2. Additionally, the visuals of the values given in Table 2 are shared in detail in Figure 2.

Variables with an equal range of values are given combined. The left y-axis reflects the range of the observed/predicted variable values, the right y-axis reflects the range of the residual values, and the x-axis reflects the order of days from 1 up to 1233. The blue lines in the graphs show the observed values, the red lines the prediction values, and the green lines the residual values. To evaluate the graphic, it is necessary to evaluate the individual and relative positions of the lines separately. It can be interpreted such that the blue and red lines get closer to each other, the model performs better. From this respect, the distribution of the green lines around the line x = 0 shows that the model is quite successful.

To check residual independence, a scatterplot is drawn according to a time variable (date). A nonrandom distribution shows that the independence condition fails. Accordingly, a scatter-plot was drawn for each variable and it was observed that the residual independence conditions are satisfied for all variables.

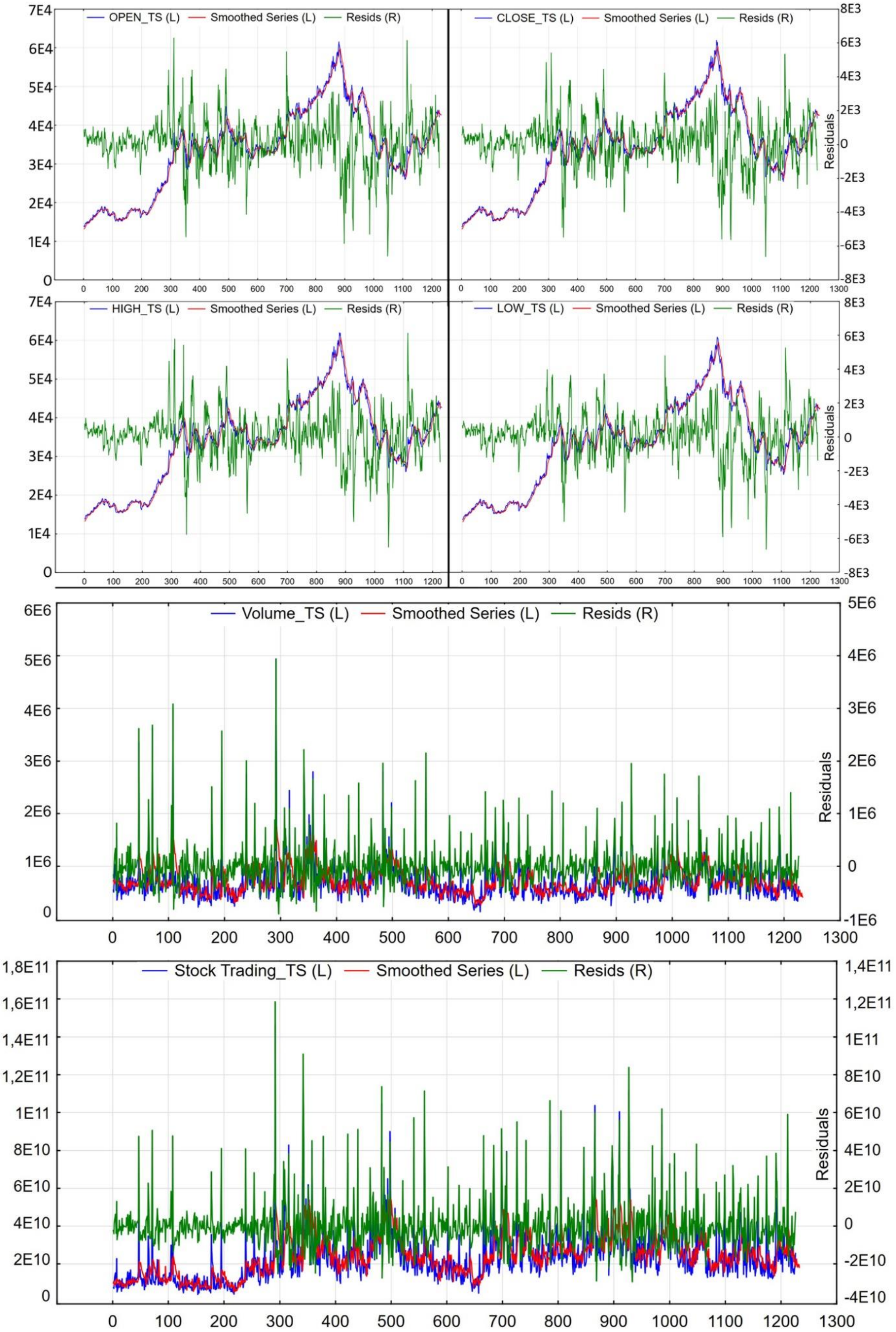

Figure 2. Time-Series prediction

In the next step, the variables generated by the time-series-smoothing model are deployed in the MARSplines model to estimate the dependent variable. This process is applied only to the training set. In this way, the prediction achievement of the model can be measured by comparison so that trials can be performed to develop a better model. However, when working on the testing set, the dependent variable is censored, and an estimation is made by applying the most successful model designed on the training set. In other words, there is no trial-development process on the testing set. Only the best available model can be used. Also, considering that the data set covers 5 years, it is highly probable that there may be seasonal effects. For this reason, the seasonal effect parameter is also included in the model. The number of seasons (periods) is determined as 12 correspondings to the 12 months in a year. However, the testing set consists of only 7 cases, and a regression model built on such a small set would not be able to train the model enough. For this reason, the mathematical infrastructure of the most successful MARSplines model created on the training set is applied to the testing set. Details of the model are presented in Table 4. Predictive Model Markup Language (PMML) of the model is given in Appendix.

Table 3. Original and predicted stock trading prices

	Dependent Variable: Stock Trading Prices		
	Observed_Stock	**Time_Series_Stock**	**MARSplines_Stock**
TESTING SET	2,79E^{+}10	2,13E^{+}10	2,33E^{+}10
	2,21E^{+}10	2,26E^{+}10	2,44E^{+}10
	5,77E^{+}10	1,95E^{+}10	2,12E^{+}10
	4,61E^{+}10	1,84E^{+}10	2,07E^{+}10
	2,11E^{+}10	1,97E^{+}10	2,31E^{+}10
	3,05E^{+}10	1,96E^{+}10	2,16E^{+}10
	5,12E+10	1,82E^{+}10	2,04E^{+}10

Table 4. MARSplines model results (significant b* are highlighted in red)

MARSplines Results	**Regression Statistics**	
Dependent: Stock Trading_TS	Mean (observed)	2,44E^{+}10
Indep.Var.: Open_TS, High_TS, Low_TS, Close_TS, Volume_TS	Std. Dev. (observed)	1,52E^{+}10
Number of terms = 13	Mean (predicted)	2,44E^{+}10
Number of basis functions = 12	Std. Dev. (predicted)	1,48E^{+}10
Order of interactions = 1	Mean (residual)	4,93E^{-}6
Penalty = 2	Std. Dev. (residual)	3,35E^{+}9
Threshold = 0,0005	R-square	9,51E^{-}1
GCV error = 1,17E^{+}19	R-square adjusted	9,51E^{-}1

In the testing set, 7 cases of the dependent variable are calculated using the function (Eq. 9) obtained from the MARSplines model. In other words, the model obtained from the training set is applied to the testing set. Stock_Trading values that are calculated according to the function are given in Table 3.

Let
$Stock\_Trading\_TS = Y, Volume\_TS = X_1, High\_TS = X_2, Low\_TS = X_3,$ and $Open\_TS = X_4$
such that the mathematical expression of the MARSplines regression model is

$$
\begin{aligned}
Y = 9{,}57E^{+}10 \ & -3{,}16E^{+}4 \cdot max(0;\ 2{,}08E^{+}6 - X_1) + 4{,}03E^{+}6 \cdot max(0;\ X_2 - 4{,}49E^{+}4) \\
& -1{,}91E^{+}6 \cdot max(0;\ 4{,}49E^{+}4 - X_2) + 2{,}15E^{+}6 \cdot max(0;\ X_3 - 4{,}30E^{+}4) \\
& +1{,}37E^{+}4 \cdot max(0;\ X_1 - 1{,}63E^{+}6) - 1{,}18E^{+}6 \cdot max(0;\ X_4 - 2{,}68E^{+}4) \\
& +1{,}33E^{+}6 \cdot max(0;\ 2{,}68E^{+}4 - X_4) - 4{,}49E^{+}6 \cdot max(0;\ X_3 - 4{,}51E^{+}4) \\
& +2{,}25E^{+}6 \cdot max(0;\ X_4 - 4{,}73E^{+}4) - 2{,}95E^{+}6 \cdot max(0;\ X_3 - 4{,}06E^{+}4) \\
& \\
& +9{,}04E^{+}5 \cdot max(0;\ X_2 - 3{,}77E^{+}4) - 7{,}54E^{+}5 \cdot max(0;\ X_3 - 5{,}48E^{+}4)
\end{aligned} \tag{9}
$$

Table 3 shares the original (observed), time-series predicted, and stock_trading values calculated using the MARSplines function. In the next step, a time series model is created on the values calculated according to the function obtained from the MARSplines model. The performance of the integrated approach of MARSplines and time-series algorithms is measured by the similarity rate between the predicted and observed values. For this purpose, first of all, the absolute correlation between predicted and observed values is compared. In the next step, the predictive success rate of the model is calculated.

Table 5. Absolute Correlation between the Stock Trading Prices (significant corr. is highlighted in red)

	Observed
Predicted (Time Series)	Abs. Corr. = 0,678 P-Value = 0,094
Predicted (Integrated Model)	Abs. Corr. = 0,852 P-Value = 0,015

The correlation rate between predicted and observed values can be used as a criterion for evaluating model performance [18]. The absolute correlation values between predicted and observed values are given in Table 5. Accordingly, the absolute correlation ratio between the predicted values based on the MARSPlines model and the observed values is 0.852. Also, the ratio of absolute correlation between predictive values based on time series and observed values is 0.678. In the table, the statistically significant correlation values are highlighted in red. According to the results, as MARSplines based correlation values are significant at α = 0.05 significance level, time-series based correlation values are not statistically significant.

Table 6. Multiple Regression Summary Statistics

Summary Statistics			
Multiple R	0,908	**F(2,4)**	9,44
Multiple R^2	0,825	**p-value**	0,031
Adjusted R^2	0,737	**Std.Err.(Estim.)**	$7{,}56E^{+}9$

Table 7. Multiple Regression Parameter Estimations (significant b* are highlighted in red)

Regression Summary (Dependent Variable: Original_Stock)						
N=7	**b***	**Std.Err. (b*)**	**b**	**Std.Err.**	**t(4)**	**P-Value**
Intercept			2,31E+11	$4{,}56E^{+}10$	5,06	0,007
TS_Stock	0,85	0,569	8,01	5,34	1,49	0,208
MARS_Stock	-1,64	0,569	-16,1	5,53	-2,89	0,044

In the second evaluation step, a multiple regression model is designed. According to the performance of the predicted sets within the model, the levels of proximity to the original data are observed. The dependent variable of the regression is ’Original_Stock and the independent variables are MARSplines_Stock and Time-Series_Stock. Summary statistics of the model are given in Table 6. Also, the parameter estimations are given in Table 7. According to the results, while the contribution of MARSplines_Stock as a predictor in the model is statistically significant, Time-Series_Stock's contribution is not significant.

Based on both evaluation steps, the predicted sets based on the MARSplines method perform better and statistically significant. These results confirm that the supporting time-series with MARSplines increases the predictive performance of the time-series.

4. Conclusion

This study proposes a new integrated prediction algorithm that functionally expresses the distribution by time-variable. In recent years, there have been numerous increases in the number of user-generated reviews with external intervention, which are quite accurate but require a deeper understanding of machine learning methods. The main contribution of the proposed approach presented in this study is to give accurate results at big sized data while reducing the outside intervention of the researcher and the amount of data required for model training. Therefore, a new approach is presented to improve the prediction performance of the time series algorithm. It enables the use of data sets from different domains with any size and number of variables, without the need for advanced understanding in machine learning. For this purpose, a dataset that is difficult to predict, nonlinear and open to irrational impact was chosen to get the best test for the performance of the approach: Stock market data. The emphasis of the experimental study is based on the examination of the data produced by the market on a continuous scale with a public license. For this purpose, the five years of the Tokyo Stock Exchange was examined. After experimental studies on the relevant data, a new algorithm has been developed to improve the current prediction potential. An exponential smoothing time series algorithm was applied to the data and predictions were made for future data. Then, another estimation was made using the MARSplines model developed with predictors based on time series. The dataset generated based on predictions obtained by MARSplines had a higher absolute correlation with the original data. (0.852-0.678). Also, MARSplines predictions are statistically significant and outperform normal time-series predictions.

References

[1] Guolo, C. Varin. (2014). Beta Regression For Time Series Analysis Of Bounded Data, With Application To Canada Google R Flu Trends, The Annals of Applied Statistics, Institute of Mathematical Statistics, Vol. 8, No. 1, 74–88, Doi: 10.1214/13-AOAS684

[2] S. Arasu, M. Jeevananthan, N. Thamaraiselvan, B. Janarthanan. (2014). Performances of data mining techniques in forecasting stock index – evidence from India and US, J.Natn.Sci.Foundation Sri Lanka 42 (2): 177–191, DOI: http://dx.doi.org/10.4038/jnsfsr.v42i2.6989

[3] Friedman, J. (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1), 1-67. Retrieved April 7, 2021, from http://www.jstor.org/stable/2241837

[4] J. R. Leathwick, D. Rowe, J. Richardson, J. Elith, T. Hastie. (2005). Using multivariate adaptive regression splines to predict the distributions of New Zealand’s freshwater diadromous fish, Freshwater Biology 50, 2034–2052, doi:10.1111/j.1365-2427.2005.01448.x

[5] M. M. Al-Idrisi. (1991). Use of Regression and Triple Exponential Smoothing Models for Forecasting Share Prices of Saudi Companies, JKAU: Econ. & Adm. vol. 4, pp. 3-25 (1411 A.H. / 1991 A.D.)

[6] M. C. A. Neto, G. Tavares, V. M. O. Alves, G. D. C. Cavalcanti, T. I. Ren. (2010). Improving financial time series prediction using exogenous series and neural networks committees, The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, 2010, pp. 1-8., doi: 10.1109/IJCNN.2010.5596911

[7] Gelper, S., Fried, R. and Croux, C. (2010), Robust forecasting with exponential and Holt–Winters smoothing. J. Forecast., 29: 285-300. https://doi.org/10.1002/for.1125

[8] McKenzie, E. (n.d.). General exponential smoothing and the equivalent arma process. Journal of Forecasting, 3(3), 333–344. https://doi.org/10.1002/for.3980030312

[9] O. Kisi, K. S. Parmar. (2016) Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J Hydrol 534:104–112]

[10] O. Hamidi, L. Tapak, H. Abbasi, Z. Maryanaji. (2017). Application of random forest time series, support vector regression and multivariate adaptive regression splines models in prediction of snowfall (a case study of Alvand in the middle Zagros, Iran). Theoretical and Applied Climatology. 134. 1-8. 10.1007/s00704-017-2300-9.

[11] L. J. Kao, C. C. Chiu. (2020). Application of integrated recurrent neural network with multivariate adaptive regression splines on SPC-EPC process, Journal of Manufacturing Systems, Vol. 57, Pages 109-118, ISSN 0278-6125, doi.org/10.1016/j.jmsy.2020.07.020.

[12] Grillenzoni, M. Fornaciari. (2019). On-line peak detection in medical time series with adaptive regression methods. Econometrics and Statistics, Vol. 10, Pages 134-150, ISSN 2452-3062, doi.org/10.1016/j.ecosta.2018.07.002.

[13] S. Lee, S. Lee, M. Moon. (2020). Hybrid change point detection for time series via support vector regression and CUSUM method. Applied Soft Computing, Vol. 89, 106101, ISSN 1568-4946, doi.org/10.1016/j.asoc.2020.106101.

[14] G. R. Mode, K. A. Hoque. (2020). Adversarial Examples in Deep Learning for Multivariate Time Series Regression. eprint=2009.11911, arXiv, cs. LG

[15] Y. Okkaoglu, Y. Akdi, , E. Golveren, M. Yucel. (2020). Estimation and forecasting of PM10 air pollution in Ankara via time series and harmonic regressions. International Journal of Environmental Science and Technology. (doi:10.1007/s13762-020-02705-0)

[16] S. Jiang. (2019)."Combining Deep Neural Networks and Classical Time Series Regression Models for Forecasting Patient Flows in Hong Kong," in IEEE Access, vol. 7, pp. 118965-118974, doi: 10.1109/ACCESS.2019.2936550.

[17] I. Ilic, B. Gorgulu, M. Cevik, M. G. Baydogan. (2020). Explainable boosted linear regression for time series forecasting. eprint=2009.09110, arXiv, cs.LG

[18] Hjort, Jan & Suomi, Juuso & Käyhkö, Jukka. (2011). Spatial prediction of urban–rural temperatures using statistical methods. Theoretical and Applied Climatology. 106. 139-152. 10.1007/s00704-011-0425-9.

Appendix

```
<?xml version="1.0" encoding="windows-1254" ?>
<PMML version="2.0">
<Header copyright="STATISTICA Data Miner, Copyright (c) 2014, StatSoft, Inc, www.StatSoft.com"/>
<MARSplinesModel>
        modelName = "Multivariate Adaptive Regression Splines"
        modelType = "MARSplines"
        <ResponseList>
        <Response name="Stock Trading_TS" noNominals="0"/>
        </ResponseList>
        <PredictorList>
                <Predictor name="Open_TS" noNominals="0"/>
                <Predictor name="High_TS" noNominals="0"/>
                <Predictor name="Low_TS" noNominals="0"/>
                <Predictor name="Close_TS" noNominals="0"/>
                <Predictor name="Volume_TS" noNominals="0"/>
</PredictorList>
<Terms>
<TermSpecs  term="1" predictor="5" knot="2,07890000000000e+006" type="-1"/>
<TermSpecs  term="2" predictor="2" knot="4,49200000000000e+004" type="1"/>
```

```
<TermSpecs  term="3" predictor="2" knot="4,49200000000000e+004" type="-1"/>
<TermSpecs  term="4" predictor="3" knot="4,30000000000000e+004" type="1"/>
<TermSpecs  term="5" predictor="5" knot="1,62610000000000e+006" type="1"/>
<TermSpecs  term="6" predictor="1" knot="2,68000000000000e+004" type="1"/>
<TermSpecs  term="7" predictor="1" knot="2,68000000000000e+004" type="-1"/>
<TermSpecs  term="8" predictor="3" knot="4,51250000000000e+004" type="1"/>
<TermSpecs  term="9" predictor="1" knot="4,72800000000000e+004" type="1"/>
<TermSpecs  term="10" predictor="3" knot="4,06100000000000e+004" type="1"/>
<TermSpecs  term="11" predictor="2" knot="3,76800000000000e+004" type="1"/>
<TermSpecs  term="12" predictor="3" knot="5,47600000000000e+004" type="1"/>
</Terms>
<ParameterList>
<Coefficients response="1" intercept="9,57354991253548e+010" coeff1="-3,16012041310131e+004"
coeff2="4,02943152348027e+006" coeff3="-1,90567698334054e+006" coeff4="2,14802239018978e+006"
coeff5="1,37012891096041e+004" coeff6="-1,18261460614068e+006" coeff7="1,33241133917770e+006"
coeff8="-4,48816386471684e+006" coeff9="2,25246971605447e+006" coeff10="-2,95446036624414e+006"
coeff11="9,03774315795363e+005" coeff12="-7,53618316284358e+005"/>
</ParameterList>
<?The following model should be used directly, with categorical variables being coded 0, 1.?>
<Stock Trading_TS = 9,57354991253548e+010 - 3,16012041310131e+004*max(0; 2,07890000000000e+006-
Volume_TS) + 4,02943152348027e+006*max(0; High_TS-4,49200000000000e+004) -
1,90567698334054e+006*max(0; 4,49200000000000e+004-High_TS) + 2,14802239018978e+006*max(0;
Low_TS-4,30000000000000e+004) + 1,37012891096041e+004*max(0; Volume_TS-1,62610000000000e+006) -
1,18261460614068e+006*max(0; Open_TS-2,68000000000000e+004) + 1,33241133917770e+006*max(0;
2,68000000000000e+004-Open_TS) - 4,48816386471684e+006*max(0; Low_TS-4,51250000000000e+004) +
2,25246971605447e+006*max(0; Open_TS-4,72800000000000e+004) - 2,95446036624414e+006*max(0;
Low_TS-4,06100000000000e+004) + 9,03774315795363e+005*max(0; High_TS-3,76800000000000e+004) -
7,53618316284358e+005*max(0; Low_TS-5,47600000000000e+004)>
</MARSplinesModel>
</PMML>
```