Heliyon 9 (2023) e19964

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Model selection in multivariate adaptive regressions splines (MARS) using alternative information criteria

Meryem Bekar Adiguzel [a,*], Mehmet Ali Cengiz [b]

[a] *Department of Finance, Banking & RInsurance, Ortakoy Vocational School of Higher Education, Aksaray University, 68400, Ortakoy, Aksaray, Turkey*

[b] *Department of Statistics, Faculty of Science and Letters Institute of Graduate Studies, Ondokuz Mayıs University, 55270, Samsun, Turkey*

ARTICLE INFO

Keywords:
Multivariate Adaptive Regression Splines
Model selection
Akaike information criterion
Schwarz Bayesian information criterion
Information complexity criterion

ABSTRACT

Multivariate Adaptive Regression Splines (MARS) is a useful non-parametric regression analysis method that can be used for model selection in high-dimensional data. Since MARS can identify and model complex, non-linear relationships between the dependent variable and independent variables without requiring any assumptions, it has advantage over simple linear regression techniques. Also, for simplifying the model building process and preventing overfitting, MARS can select automatically the variables to be included in the model, which is useful for datasets with many variables. While MARS is a flexible non-parametric regression method, generalized cross validation (GCV) technique is used within the MARS framework to avoid overfitting and to select the best model. GCV criterion is widely used and can be effective in many situations, however it has some criticism. These criticism are the arbitrary value of the smoothing parameter used in the algorithm of the GCV criterion and the models obtained using this criterion are high-dimensional. In this paper, it is aimed to obtain the barest model that best explains the relationship between the dependent variable and independent variables by using alternative information criteria (Akaike information criterion (AIC), Schwarz Bayesian criterion (SBC) and information complexity criterion ($\mathrm{ICOMP(IFIM)_{PEU}}$)) instead of the use of smoothing parameters in order to put an end to the criticism. To achieve this goal, a simulation study was first conducted with a data set composed of variables that do and do not contribute to the dependent variable to test the success of the information criteria. As a consequence of this simulation work, when variables (which do not contribute to the dependent variable) are not included in the regression model, it demonstrates the success of the criteria in model selection. As a real data set, the reasons for loan defaults were investigated between the years 2005–2019 by utilizing data from 18 banks operating in Türkiye. The results obtained reveal the success of $\mathrm{ICOMP(IFIM)_{PEU}}$ criterion in model selection.

1. Introduction

Multivariate Adaptive Regression Splines (MARS) is a non-parametric regression analysis method. In other words, it makes less assumptions about the form of the underlying relation between the independent variables and the dependent variable. Flexibility, interactions, piecewise linearity and model selection are some key characteristics of the MARS technique. By means of model selection

[*] Corresponding author.
E-mail address: bekarmeryem@gmail.com (M. Bekar Adiguzel).

https://doi.org/10.1016/j.heliyon.2023.e19964
Received 16 April 2023; Received in revised form 22 August 2023; Accepted 7 September 2023
Available online 17 September 2023

MARS uses a two-step process (a forward step and a backward step) to build the model based on certain criteria like generalized cross validation (GCV) [1]. However, it has some certain limitations that can attract criticism. The first criticism is the arbitrary value of the smoothing parameter used in the algorithm of the GCV criterion, while the second one is the models obtained using this criterion are high-dimensional [2]. Despite the criticism, MARS has been used as a successful model selection tool as a non-parametric modeling technique in many fields of science. For example; effect of gender on musculoskeletal disorders [3]; possible impact of climate change on temperature and precipitation parameters in the Eastern Black Sea region [4]; how the use of traditional media and social media effect the political attitudes and behaviors of citizens [5]; worldwide cases of rockburst due to severe damage to infrastructures and facilities, and factors triggering these rockburst and the rockburst intensity [6] and comparison of the effectiveness of Islamic bank in developed and developing countries and the relationship between the efficiency of these banks with gross domestic product (GDP) and Sharia Supervisory Board [7] are examined using MARS technique. Model selection performances of the MARS method (which is obtained by using the GCV criterion) and the Random Ensemble MARS (REMARS) method (which is obtained by using the random forest algorithm) are compared. For more reliable results, MARS model is recommended for large datasets, while REMARS model is recommended for small datasets [8].

Some of the studies on information criteria in the literature are as follows: the relationship of obesity with age, height, weight and different parts of the body is investigated by using information complexity (ICOMP) criterion as the fitness function in the MARS technique [2]. It has been claimed that the inverse of the Gram matrix of the Elastic Net (EN) modelling method can not be calculated, so this method cannot be used for model selection of undersized samples with high-dimensions. To overcome this problem, a new adaptive elastic net (AEN) approach is proposed using the ICOMP criterion [9]. The performances of Akaike information criterion (AIC) and ICOMP criteria are compared. As a result of the comparison, ICOMP criterion based on M, S and MM estimators (The S and MM estimations are improved versions of the M estimation based on the maximum likelihood method [10]) is being proposed [11]. The relationship between energy consumption and the amount of CO_2 is examined by comparing the AIC, Akaike information criterion corrected (AICC), Schwarz Bayesian criterion (SBC) and Hannan-Quin information criterion (HQC) [12]. Since criteria such as AIC and Bayesian information criterion (BIC) have a tendency to overfit in model selection in high-dimensional data; EBIC (extended BIC), EBIC-Robust and EFIC (it is formed by combining extended fisher information criterion and EBIC) have been proposed instead of these criteria. The results showed that the EBIC-Robust criterion performed better in model selection compared to the EFIC and EBIC criteria [13].

The main purpose of this study is to find the barest model that explains the dependent variable in the MARS technique by using alternative information criteria (AIC, SBC, $\text{ICOMP}(\text{IFIM})_{\text{PEU}}$) instead of the GCV criterion and to put an end to the criticism about the arbitrary value of the smoothing parameter. Firstly, we introduce the MARS technique and GCV criterion, then we give the definitions of AIC, SBC and ($\text{ICOMP}(\text{IFIM})_{\text{PEU}}$) criteria, respectively. Finally, the success of the information criteria in model selection is examined with a simulation study and the study is continued on a real dataset. The dataset used in this study is based on the real data of 18 banks operating in Turkey for a 15 year period, 2005–2019. The reasons of non-performing loans (NPL) are being examined by using this dataset. Compatible with the relevant literature, in this study NPL ratio is taken as the dependent variable while inflation, industry production index, exchange rate, unemployment, interest rate, imports, exports and gross domestic product (GDP) are taken as the independent variables.

2. Multivariate adaptive regression splines

MARS, developed by Friedman in 1991, is a multivariate non-parametric regression. Using 'smoothing splines' technique, the MARS method provides both innovation and a great convenience in terms of interpreting the variables and their interactions [1]. MARS, which can estimate models of high dimensional datasets, is a powerful and useful method. The form of MARS is

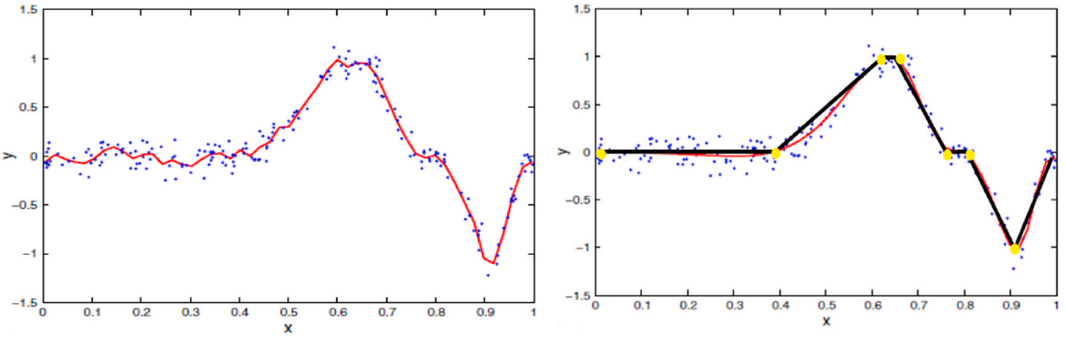
$$Y = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x) + \varepsilon$$

where $x = (x_1, x_2, \ldots, x_p)^T$, the error term ε is normally distributed with zero expected value and constant variance M is the number of the basis functions (BFs), $B_m(x)$ is the m th BF, β_0 is the coefficient for the constant, and β_m is the coefficient for the m th BF. The m th BF form is as follows

$$B_m(x) = \prod_{k=1}^{K_m} \left[S_{km} \cdot \left(x_{v(k,m)} - t_{k,m} \right) \right]_+$$

Here; K_m represents the number of truncated linear functions, that is, multiplied in the m th each basis function $x_{v(k,m)}$ is the vector of independet variable, $t_{k,m}$ is the knot value (breakpoint), and $S_{km} = \pm 1$.

MARS method performs model selection in two stages. First stage is the forward selection stage in which all possible basis functions are obtained and thus the model with maximum complexity is formed. The second stage is the backward elimination stage, where the complexity of the model is reduced [1,14]. At this stage, the basis function that causes an increase in the residual sum of squares are eliminated at each process. This process continues until the GCV criterion reaches its minimum value for the overall model [15]. The MARS technique uses GCV criterion for model selection [16]. GCV criterion is used as a measure for the degree of fit of accuracy of the model [17]. Additionally, using GCV in MARS, then it yields to the lack-of-fit (LOF) of the model [14].



(a)

(b)

Fig. 1. An example indicating the formation phase of the MARS model
(a) Modelling with complex regression model, (b) MARS modelling with truncated linear function.

$$LOF(\widehat{f}_M) = GCV(M) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \widehat{f}_M(x_i))^2}{\left(1 - \frac{P(M)^*}{n}\right)^2} \tag{1}$$

$$P(M) = trace\left(\boldsymbol{B}\left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}\boldsymbol{B}^T\right) + 1 \tag{2}$$

$$P(M)^* = P(M) + dM$$

GCV is associated with the idea of minimizing residual sum of squares [18]. In (1); y_i is the *i-th* observed response variable vector, $\widehat{f}_M$ is the MARS estimate model based on the basis function, $\mathrm{x_i}$ is the *i-th* observed independent variable vector, $x_i = (x_{i1}, \ldots, x_{ip})^T, i = (1, \ldots, n)$, the number of observations in the dataset is denoted by *n*, the number of BFs in the model is denoted by *M*, the matrix of BFs with dimension *Mxn* is represented by *B*. $P(M)$ in (2) represents the penalty measure for complexity, and d is defined as the smoothing parameter. A large value of d means obtaining a barest model estimation (i.e., resulting in fewer BFs) [1,2]. As the model is pruned, eliminated values are called as penalty. Studies in recent years show that for d, the best value is in the range $2 \le d \le 4$ [1,19]. The best MARS model is the one that has the minimum GCV measurement [1,2,19]. However, there are some aspects of the GCV criterion that can attract criticism. These criticism are that the smoothing parameter used in the algorithm of the GCV criterion is arbitrary value and the models obtained using this criterion are high-dimensional [2].

The more number of the variable data, the harder to get and interprete the regression equation. MARS technique is used to make the model which has the same regression structure as in the first figure to make it more understandable. MARS method makes the quadratic or cubic model formed in the forward step free from this complex structure and makes it piecewise linear see Fig. 1 The aim is to get the barest model that has the highest information that balances the overfitting and lack-of-fit of the model.

3. Information criteria

Considering the terms "model selection (i.e., finding the barest model that 'best explains the dependent variable', 'consisting of independent variables' and 'contains the most information')" and "Occam's Razor or law of parsimony (i.e., the simpler one is considered to be better in the presence of two competing theories making the same predictions)", it is very important to take account the complexity of the model. Thus, in model selection it is aimed to find the barest model by using the alternative informative criteria (AIC, SBC and ICOMP(IFIM)$_{\mathrm{PEU}}$) given below.

3.1. Akaike information criterion (AIC)

AIC was suggested by Akaike in 1973 as "minus two times the log likelihood plus two times the number of model parameters":

$$AIC = -2logL(\widehat{Q}) + 2k$$

where $L(\cdot)$ represents the likelihood function, $\widehat{Q}$ is the maximized likelihood (ML) estimate of parameter vector Q and k is the number of the independent parameters [20,21]. The term $-2logL(\widehat{Q})$ measures the model fit. The more information about the model, the larger the value $2logL(\widehat{Q})$, but because of the minus sign the value of the first term $-2logL(\widehat{Q})$ will be smaller [1]. As it is well known, as the number of variables or parameters in a model increase, the model is explained more better and its variances get smaller. It is not correct to add parameters to the model just to increase the explanation percentage of the model. This situation does not coincide with the concept of finding the best model using the least number of parameters in the model selection definition. For this reason, in order to prevent the unnecessary increase in the number of parameters, the information criteria includes a penalty term that increases the value of the criteria as the number of parameters increase. Thus, when an unnecessary parameter is added to the model, the penalty term will increase the value of the information criterion and make the selection of the model difficult.

3.2. Schwarz Bayes information criterion (SBC)

SBC, has been suggested by Schwartz in 1978 and is closely releated to the AIC. The formula of SBC differs only from that of the AIC in the second term; SBC's penalty term $klog(n)$ is greater than AIC's $2k$. So, the model selected with SBC is expected to the smaller or equal in size than the model selected with AIC [22].

$$SBC = -2logL(\widehat{Q}) + klog(n)$$

3.3. $ICOMP(IFIM)_{PEU}$ criterion

ICOMP is a criterion defined by Bozdoğan in 1988 [23–27]. ICOMP uses the covariance complexity of the model as a penalty term instead of the number of free parameters. It aims to find the best balance between overfitting and lack-of-fit of the model. Moreover, it does this by finding the barest model that contains the most information [23]. The general form of ICOMP is

$$ICOMP = -2logL(\widehat{Q}) + 2C_1(\Sigma) \tag{3}$$

where $L(\cdot)$ represents the likelihood function, $\widehat{Q}$ is the maximized likelihood (ML) estimate of parameter vector Q, C_1 corresponds to the complexity measure and finally Σ represents the estimated covariance matrix of the parameter vector. Using the inverse Fisher information matrix

$$\widehat{F}^{-1} = \left\{ -E\left(\frac{\partial^2 logL(\mathrm{Q})}{\partial \mathrm{Q} \partial \mathrm{Q}'}\right)_{\hat{\mathrm{Q}}} \right\}^{-1}$$

instead of the covariance estimation in (3), we obtain

$$ICOMP(IFIM) = -2logL(\widehat{Q}) + 2C_1\left(\widehat{F}^{-1}\right)$$

where

$$C_1\left(\widehat{F}^{-1}\right) = (M+1)log\left[\frac{tr\widehat{\sigma}^2(B'B)^{-1} + \frac{\widehat{2\sigma^4}}{n}}{M+1}\right] - log\left|\widehat{\sigma}^2(B'B)^{-1}\right| \, log\frac{2\widehat{\sigma}^4}{n}$$

and

$$Cov\left(\widehat{\beta}, \widehat{\sigma}^2\right) = \widehat{F}^{-1} = \begin{bmatrix} \widehat{\sigma}^2(B'B)^{-1} & 0 \\ 0 & \frac{2\widehat{\sigma}^4}{n} \end{bmatrix}$$

here, M shows the number of BFs in model while B shows the number of free parameters.

The ICOMP(IFIM) criterion is improved as a Bayesian criterion by maximizing the posterior expected utility (PEU), which is given by

$$ICOMP(IFIM)_{PEU} = -2logL(\widehat{Q}) + k(1 + log(n)) + 2C_1\left(\widehat{F}^{-1}\right)$$

that provides severe penalization for the excessive parameterization with the $k(\mathbf{1} + \boldsymbol{log}(\boldsymbol{n}))$ term [25,27]. And thus, a simple MARS model than AIC and SBC will be obtained [2].

4. Simulation study

A simulation study is carried out to test the success of the information criteria used in MARS in model selection. For the simulation study the program RStudio is used. Firstly, 10 independent variables containing 100 observations are generated from a uniform distribution between 0 and 1. The first 5 variables are selected as contributing variables to the model, and the last 5 as non-contributing variables. So the correct model would be expected to contain the first five variables x_1, x_2, x_3, x_4 and x_5. This selection will show the success of the information criteria in model selection.

Afterwards, the knot point for each variable is assigned as 3 [1]. In other words, knot points will be occurred in each 33 data. Thus, for each variable there occurs 3 BFs. Error variable vector is also generated by using standard normal distribution $\varepsilon \sim N(0,1)$. Finally, the vector y is constructed with the initial equation $y = 10\, sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10\, x_4 + 5\, x_5 + 0.5\, \varepsilon$ [2]. In this way, x matrix

Table 1
Representation of BFs for the new dataset with 1830 independent variables and 100 observations.

$x_{1,1} - knot_{x1,1}$	$x_{1,2} - knot_{x1,2}$	$\cdots$	$(x_{1,10} - knot_{x10,3})(knot_{x10,3} - x_{1,10})$
$x_{2,1} - knot_{x1,1}$	$x_{2,2} - knot_{x1,2}$	$\cdots$	$(x_{2,10} - knot_{x10,3})(knot_{x10,3} - x_{2,10})$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{100,1} - knot_{x1,1}$	$x_{100,2} - knot_{x1,2}$	$\cdots$	$(x_{100,10} - knot_{x10,3})(knot_{x10,3} - x_{100,10})$

and y, ε vectors are obtained. 30 BFs, 3 from each variable, are obtained and 60 BFs are formed with mirror functions. Another advantage of MARS is taking the variables interaction to the effect of the model into consideration. When the interactions of 60 basis functions {BF1BF2, …BF1BF60, BF2BF3, …BF2BF60, BF3BF4, …BF3BF60, …BF58BF60, BF59BF60} is created, a total of 1770 BFs containing binary interactions will be obtained. 1831 variables are created, 60 of which are BFs with main effect (non-interacting), 1770 BFs with interaction and y vectors. Thus, our data set has turned into high-dimensional data (see Table 1).where $x_{i,j}$ is the i-th observation value of the variable x_j for $i = 1, \ldots, 100$ and $j = 1, \ldots, 10$, $knot_{x_{j,t}}$ is the t-th knot value obtained from the variable x_j $t = 1,2, 3$. Note that the first column above corresponds to BF$_1$, second column corresponds to BF$_2$ and the 1830-th column corresponds to BF$_{59}$BF$_{60}$. With the new variables (1830 new BFs variables will be used instead of the previous 10 X variables), the MARS model will be created. At this stage, the barest model is founded by using the alternative information criteria instead of the GCV criterion used by the MARS technique in model selection. The MARS models obtained using GCV and other information criteria, and the ANOVA decomposition table obtained using GCV and ICOMP(IFIM)$_{\text{PEU}}$ criteria are given below:

MARS model obtained by GCV criterion:

$$Y = 2.41 + 2.31\text{BF4} + 3.29\text{BF7} + 9.79\text{BF11} + 3.06\text{BF13} - 0.33\text{BF31} + 3.88\text{BF38} + 2.36\text{BF39} + 2.34\text{BF1BF5} - 11.44\text{BF1BF6} + 4.3\text{BF1BF34}$$

$$BF4 = max\ (0, x_2 - 0.179)\ BF7 = max\ (0, x_3 - 0.6578)$$

$$BF11 = max(0, x_4 - 0.0158)\ BF13 = max(0, x_5 - 0.0255)$$

$$BF31 = max(0, 0.4922 - x_1)\ BF38 = max(0, 0.6367 - x_3)$$

$$BF39 = max(0, 0.1594 - x_3)\ BF1BF5 = max(0, (x_1 - 0.4922)(x_2 - 0.674))$$

$$BF1BF6 = max(0, (x_1 - 0.4922)(x_2 - 0.0832))\ BF1BF34 = max(0, (x_1 - 0.4922)(0.179 - x_2))$$

The MARS model obtained using the GCV criterion contains 10 BFs. Three of them are included in the model as interaction functions. Interaction variables are x_1x_2. We can say that GCV criterian's success is high in model selection as fitted model doesn't contain x_6, x_7, x_8, x_9, x_{10} variables.

The ANOVA decomposition table of the MARS model obtained with the GCV criterion is given in Table 2. In this table, general effects of each variable of the final model on the dependent variables can be seen. The first column gives the number of functions. The function number specifies how many different variables the model consist of. The second column gives the standard deviation value (denoted by Std). The size of the standard deviation value shows the effect of the related variable on the dependent variable. The generalized cross validation (denoted by GCV) values in the third column shows the loss that will occur in the estimations when the variables corresponding to these values are removed from the model. The fourth column gives the number of basis functions (denoted by BFs) associated with the related variable, and the fifth column gives the number of linear degrees-of-freedom (denoted by D.f.). For example; the variable x_3 entered the model with a total of 3 basis functions, and when its standard deviation is examined, it is seen that it is the most effective variable on the dependent variable. In addition, when we look at the GCV value, we see that the loss in the model estimation will be the largest when the x_3 variable is removed from the model. As another example, when we examine the x_1 variable, we see that it has the smallest standard deviation value. Hence this variable can be removed from the model. The GCV value of the x_1 variable shows that when this variable is removed from the model, the loss will be small compared to other variables. After the variable x_3, the variables x_2, x_5 and x_1x_2, respectively, contributed the most to the model, while the variables x_1 and x_4 contributed the least to the model to the other variables.

Table 2
ANOVA decomposition that is obtained by using GCV criterion.

Function	Std	GCV	BF	D.f.	Variables
1	3,8909	15,6059	1	2,5	x_1
2	4,4838	20,7241	1	2,5	x_2
3	4,5239	25,0146	3	7,5	x_3
4	3,9784	16,3159	1	2,5	x_4
5	4,4386	20,3087	1	2,5	x_5
6	4,0955	20,5012	3	7,5	x_1x_2

MARS model obtained by AIC criterion:

$$\begin{aligned}\mathrm{Y} = & -2.65 + 2.12\mathrm{BF1} - 2.64\mathrm{BF2} + 12.16\mathrm{BF4} + 6.45\mathrm{BF5} + 0.16\mathrm{BF7} + 8.1\mathrm{BF11} + 5.54\mathrm{BF13} - 0.89\mathrm{BF20} - 8.85\mathrm{BF31} - 0.76\mathrm{BF34} \\ & - 8.03\mathrm{BF37} + 7.55\mathrm{BF38} - 14.39\mathrm{BF39} + 7.94\mathrm{BF46} - 2.61\mathrm{BF48} + 14.03\mathrm{BF49} + 8.76\mathrm{BF1BF5} + 1.6\mathrm{BF1BF6} + 3.42\mathrm{BF1BF10} \\ & - 6.32\mathrm{BF1BF28} - 2.4\mathrm{BF1BF34}\end{aligned}$$

$BF1 = max(0, x_1 - 0.4922)$ $BF2 = max(0, x_1 - 0.6096)$

$BF4 = max(0, x_2 - 0.179)$ $BF5 = max(0, x_2 - 0.674)$

$BF7 = max(0, x_3 - 0.6578)$ $BF11 = max(0, x_4 - 0.0158)$

$BF13 = max(0, x_5 - 0.0255)$ $BF20 = max(0, x_7 - 0.1337)$

$BF31 = max(0, 0.4922 - x_1)$ $BF34 = max(0, 0.179 - x_2)$

$BF37 = max(0, 0.6578 - x_3)$ $BF38 = max(0, 0.6367 - x_3)$

$BF39 = max(0, 0.1594 - x_3)$ $BF46 = max(0, 0.9516 - x_6)$

$BF48 = max(0, 0.035 - x_6)$ $BF49 = max(0, 0.7125 - x_7)$

$BF1BF5 = max(0, (x_1 - 0.4922)(x_2 - 0.674))$ $BF1BF6 = max(0, (x_1 - 0.49)(x_2 - 0.0832))$

$BF1BF10 = max(0, (x_1 - 0.4922)(x_4 - 0.5186))$ $BF1BF28 = max(0, (x_1 - 0.4922)(x_{10} - 0.7223))$

$BF1BF34 = max(0, (x_1 - 0.4922)(0.179 - x_2))$

The MARS model obtained using the AIC criterion contains 21 basis functions. The interaction variables are x_1x_2, x_1x_4 and x_1x_{10}. It is undesirable for our model to include the x_6, x_7 and x_{10} variables to select the correct model. In addition, we see that a model with quite a lot of BFs has been formed compared to the model obtained from the GCV criterion. Considering the choice of the barest model with the correct variable selection, we can say that the GCV criterion is more successful than the AIC criterion.

MARS model obtained by SBC criterion:

$$\begin{aligned}\mathrm{Y} = & -0.46 + 3.07\mathrm{BF4} + 3.06\mathrm{BF7} + 5.82\mathrm{BF11} - 0.11\mathrm{BF13} - 7.15\mathrm{BF31} - 0.38\mathrm{BF34} + 0.97\mathrm{BF37} + 3.82\mathrm{BF38} - 3.12\mathrm{BF39} \\ & + 3.14\mathrm{BF1BF5} + 8.08\mathrm{BF1BF6} + 6.92\mathrm{BF1BF34}\end{aligned}$$

$BF4 = max(0, x_2 - 0.179)$ $BF7 = max(0, x_3 - 0.6578)$

$BF11 = max(0, x_4 - 0.0158)$ $BF13 = max(0, x_5 - 0.0255)$

$BF31 = max(0, 0.4922 - x_1)$ $BF34 = max(0, 0.179 - x_2)$

$BF37 = max(0, 0.6578 - x_3)$ $BF38 = max(0, 0.6367 - x_3)$

$BF39 = max(0, 0.1594 - x_3)$ $BF1BF5 = max(0, (x_1 - 0.4922)(x_2 - 0.674))$

$BF1BF6 = max(0, (x_1 - 0.4922)(x_2 - 0.0832))$ $BF1BF34 = \max\ (0, (x_1 - 0.4922)(0.179 - x_2))$

A model containing 12 BFs is obtained by SBC criterion. In this criterion, x_1x_2 variables are determined as interacting variables. The fact that it does not include variables that do not contribute to the model show that the criterion is successful in model selection. Although SBC is not much more than the GCV criterion in terms the number of BFs, we can say that GCV has a simpler model choice.

MARS model obtained by $ICOMP(IFIM)_{PEU}$ criterion:

$$\mathrm{Y} = 1.97 + 10.01\mathrm{BF4} - 2.48\mathrm{BF5} - 1.41\mathrm{BF7} + 9.17\mathrm{BF11} - 4.78\mathrm{BF13} + 1.58\mathrm{BF31} + 4.92\mathrm{BF38} + 2.73\mathrm{BF39} + 0.42\mathrm{BF1BF34}$$

$BF4 = max(0, x_2 - 0.179)$ $BF5 = max(0, x_2 - 0.674)$

$BF7 = max(0, x_3 - 0.6578)$ $BF11 = max(0, x_4 - 0.0158)$

$BF13 = max(0, x_5 - 0.0255)$ $BF31 = max(0, 0.4922 - x_1)$

$BF38 = max(0, 0.6367 - x_3)$ $BF39 = max(0, 0.1594 - x_3)$

Tablo 3
ANOVA decomposition obtained by ICOMP(IFIM)$_{\mathrm{PEU}}$ criterion.

Function	Std.	ICOMP(IFIM)$_{\mathrm{PEU}}$	BF	D.f.	Variables
1	3,890933	560,2431	1	2,5	x_1
2	4,462392	599,0788	2	5	x_2
3	4,523992	613,3811	3	7,5	x_3
4	3,97846	566,4229	1	2,5	x_4
5	4,438636	590,2337	1	2,5	x_5
6	4,325727	589,1087	1	2,5	x_1x_2

$$BF1BF34 = \max\left(0, (x_1 - 0.4922)(0.179 - x_2)\right)$$

As it can be observed that ICOMP(IFIM)$_{\mathrm{PEU}}$ is the most successful criterion in model selection, since the ICOMP(IFIM)$_{\mathrm{PEU}}$ criterion contains 9 BFs and the variables it contains are the variables that contribute to the model.

ANOVA decomposition results obtained based on the ICOMP(IFIM)$_{\mathrm{PEU}}$ criterion of the MARS model are given in Table 3. The variable x_3 entered the model with a total of 3 BFs and when its standard deviation is examined, it is seen that it is the most effective variable on the dependent variable with a score of 4,523992. Contributing to the model, x_3 variable is followed by x_2, x_5 and x_1x_2 variables, respectively. When the effects of the variables on the model are examined, similar results are obtained in the ANOVA decomposition tables obtained using the ICOMP(IFIM)$_{\mathrm{PEU}}$ and GCV criteria.

5. Real dataset practice

The fact that banks, one of the important actors of the modern economy, can perform their functions rationally and have a sustainable infrastructure is possible with the combination of many different factors. One of these factors is the repayment of loans given by banks at a significant rate. Otherwise, the NPL ratio will increase and this will naturally affect the performance of banks negatively. There are many indicators that have the potential to affect the NPL ratio. It is possible to make different choices among these indicators. Considering the relevant literature studies, it can be said that the factors of industry production index, unemployment, exchange rate, interest rate, inflation, gross domestic product (GDP), imports and exports have a significant effect on the NPL ratio of loans. In this study, using the data of 18 banks operating in Turkey for the years 2005–2019, the relationship between the NPL ratio and the variables mentioned above will be given. While creating the data set, the NPL ratio is taken as the dependent variable, while the industry production index, unemployment, exchange rate, inflation, interest rate, imports, exports and GDP are taken as the independent variables, see Table 4.

Application is done in Rstudio program (for Rstudio program see Ref. [28]). The least squares estimator (LSE) of β coefficients vector of multiple regression model $Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$ $(i = 1,2, \ldots,n)$ is obtained by the formula $\widehat{\beta} = (X'X)^{-1}X'Y$. However, since the numerical values of the GDP, exports and imports data are very large, the natural logarithms of these data are taken and these transformed data are used in the analysis. If the part $(X'X)^{-1}$ of the LSE estimator is calculated with the data without logarithm, the β coefficients are estimated as extremely small values. This situation, which also creates difficulties in the interpretation of the regression equation, is solved by taking the logarithm of the values.

Knot values of the variables are obtained from the MARS program (for MARS program see Ref. [29]). The MARS program determines knot values with a search procedure algorithm. The MARS method starts the model selection by removing the variables that do not have knot values from the data set [30]. It has been seen that no knot value has been created for the GDP variable, and this

Table 4
List of variables for real dataset.

y: Non-performing loans of credit	
x_1 : Industry production index	x_5 : Interest rate
x_2: Unemployment	x_6 : Imports
x_3 : Exchange rate	x_7 : Exports
x_4 : Inflation	x_8 : GDP

Table 5
Knots and ranks of real data variables.

Variable	Knot		Value Without Ln		Rank	
Unemployment rate	7,3	10,1	–	–	31	16
Inflation	3,99	–	–	–	36	–
Interest rate	8,53	–	–	–	36	–
Industry production index	53,3	82,99	–	–	60	30
Exchange rate	1,2	2,26	–	–	48	21
Exports (Ln)	16,71	17,33	18.214.793	33.797.332	3	28
Imports (Ln)	17,14	17,86	27.822.542	57.234.288	1	50

variable has been removed from the data set by MARS. The knot values obtained for each variable and the ordinal numbers corresponding to these knot values are given in Table 5 as follows.

MARS model obtained by GCV criterion:

$$\begin{aligned}
\mathrm{Y} = {} & 0.0007971 + 0.0007471\mathrm{BF1} - 0.0015716\mathrm{BF2} + 0.0005623\mathrm{BF3} - 0.0001953\mathrm{BF4} + 0.0014005\mathrm{BF5} - 0.0021203\mathrm{BF6} \\
& + 0.0002353\mathrm{BF8} + 0.0000713\mathrm{BF9} - 0.0001196\mathrm{BF11} + 0.0007151\mathrm{BF13} + 0.0000815\mathrm{BF15} - 0.0008827\mathrm{BF18} + 0.0005647\mathrm{BF20} \\
& - 0.000047\mathrm{BF21} + 0.0003499\mathrm{BF1BF4} - 0.0000503\mathrm{BF1BF5} - 0.000032\mathrm{BF1BF7} + 0.0006535\mathrm{BF1BF8} - 0.0001343\mathrm{BF1BF10} \\
& - 0.0002163\mathrm{BF1BF12} - 0.0001099\mathrm{BF1BF17} + 0.0007017\mathrm{BF1BF18} - 0.0007195\mathrm{BF1BF20} - 0.0002283\mathrm{BF1BF21} \\
& + 0.0006874\mathrm{BF1BF22}
\end{aligned}$$

$$BF1 = UNEMPLOYMENT\ RATE - 7.3$$

$$BF2 = UNEMPLOYMENT\ RATE - 10.1$$

$$BF3 = INFLATION - 3.99$$

$$BF4 = INTEREST\ RATE - 8.5392$$

$$BF5 = INDUSTRY\ PRODUCTION\ INDEX - 82.9933$$

$$BF6 = INDUSTRY\ PRODUCTION\ INDEX - 53.3$$

$$BF8 = EXCHANGE\ RATE - 2.26$$

$$BF9 = \ln \_EXPORTS - 16.71$$

$$BF11 = \ln \_IMPORTS - 17.8627$$

$$BF13 = 7.3 - UNEMPLOYMENT\ RATE$$

$$BF15 = 3.99 - INFLATION$$

$$BF18 = 53.3 - INDUSTRY\ PRODUCTION\ INDEX$$

$$BF20 = 2.26 - EXCHANGE\ RATE$$

$$BF21 = 16.71 - \ln \_EXPORTS$$

$$BF1BF4 = (UNEMPLOYMENT\ RATE - 7.3)(INTEREST\ RATE - 8.5392)$$

$$BF1BF5 = (UNEMPLOYMENT\ RATE - 7.3)(INDUSTRY\ PRODUCTION\ INDEX - 82.9933)$$

$$BF1BF7 = (UNEMPLOYMENT\ RATE - 7.3)(EXCHANGE\ RATE - 1.2)$$

$$BF1BF8 = (UNEMPLOYMENT\ RATE - 7.3)(EXCHANGE\ RATE - 2.26)$$

$$BF1BF10 = (UNEMPLOYMENT\ RATE - 7.3)(\ln \_EXPORTS - 17.3359)$$

$$BF1BF12 = (UNEMPLOYMENT\ RATE - 7.3)(\ln \_IMPORTS - 17.14)$$

$$BF1BF17 = (UNEMPLOYMENT\ RATE - 7.3)(82.9933 - INDUSTRY\ PRODUCTION\ INDEX)$$

$$BF1BF18 = (UNEMPLOYMENT\ RATE - 7.3)(53.3 - INDUSTRY\ PRODUCTION\ INDEX)$$

$$BF1BF20 = (UNEMPLOYMENT\ RATE - 7.3)(2.26 - EXCHANGE\ RATE)$$

$$BF1BF21 = (UNEMPLOYMENT\ RATE - 7.3)(16.71 - \ln \_EXPORTS)$$

$$BF1BF22 = (UNEMPLOYMENT\ RATE - 7.3)(17.3359 - \ln \_EXPORTS)$$

The MARS model obtained using the GCV criterion contains 25 BFs. As it can be seen from Table 6, the model includes 12 variables. Unemployment rate, inflation, interest, industry production index, exports, imports and exchange rate variables contribute to the MARS model as the main effects. As a result we can say that the basis functions BF1, BF2 and BF13, which show the unemployment

Table 6
ANOVA decomposition obtained by using GCV criterion.

Function	Std	GCV	BF	D.f.	Variables
1	0.010177	0.00016599	3	7.5	Unemployment
2	0.011589	0.00018278	2	5.0	Inflation
3	0.011331	0.00015024	1	2.5	Interest
4	0.009101	0.00013274	3	7.5	Industry production index
5	0.011466	0.00017893	2	5.0	Exchange rate
6	0.010393	0.00014700	2	5.0	Exports
7	0.011366	0.00015118	1	2.5	Imports
8	0.009908	0.00012853	1	2.5	Unemployment-Interest
9	0.007044	0.00007952	3	7.5	Unemployment-Industry production index
10	0.007392	0.00008757	3	7.5	Unemployment-Exchange rate
11	0.006905	0.00007642	3	7.5	Unemployment-Exports
12	0.010707	0.00015602	1	2.5	Unemployment-Imports

rate, have a positive contribution to the model for the break point of 7.3, a negative contribution to the model for the break point of 10.1, and a positive contribution to the model for the break point (which enters the model as a reflected pair) of 7.3, respectively.

In Table 6, the ANOVA decomposition results of the significant variables of the MARS regression obtained depending on the GCV criterion are given. In this table, the general effects of each variable entering the final model on the dependent variable can be seen. The size of the standard deviation of a variable in the ANOVA table shows the size of the effect of that variable on the dependent variable. The GCV values in the third column show the loss that will occur in the estimations when the variables corresponding to these values are removed from the model. The fourth column gives the number of BFs of the related variable, the fifth column gives the number of linear degrees-of-freedom (denoted by D.f.). In this case, when the standard deviation and GCV values of inflation are examined, we can say that it is the most significant variable in the model. After inflation, the most important variable of the model is the exchange rate. For the degrees of freedom values in the fifth column, article [2] is taken as reference. According to this article, for each basis functions which entered to the model, the degree of freedom is determined as 2.5.

MARS model obtained with AIC criterian:

$$\begin{aligned} \mathrm{Y} = {} & 0.0002469 - 0.00059\mathrm{BF2} + 0.0007217\mathrm{BF3} - 0.0002371\mathrm{BF5} - 0.0004872\mathrm{BF6} + 0.0014673\mathrm{BF8} + 0.0008853\mathrm{BF11} \\ & - 0.000493\mathrm{BF18} + 0.0015257\mathrm{BF20} - 0.0032715\mathrm{BF21} - 0.000517\mathrm{BF1BF4} - 0.0000766\mathrm{BF1BF5} + 0.0007414\mathrm{BF1BF7} \\ & - 0.0004561\mathrm{BF1BF10} + 0.0007447\mathrm{BF1BF12} + 0.0022892\mathrm{BF1BF17} - 0.0020518\mathrm{BF1BF18} - 0.0019613\mathrm{BF1BF20} \\ & + 0.001582\mathrm{BF1BF21} + 0.0008637\mathrm{BF1BF22} \end{aligned}$$

$$BF2 = UNEMPLOYMENT\ RATE - 10.1$$

$$BF3 = INFLATION - 3.99$$

$$BF5 = INDUSTRY\ PRODUCTION\ INDEX - 82.9933$$

$$BF6 = INDUSTRY\ PRODUCTION\ INDEX - 53.3$$

$$BF8 = EXCHANGE\ RATE - 2.26$$

$$BF11 = ln\_IMPORTS - 17.8627$$

$$BF18 = 53.3 - INDUSTRY\ PRODUCTION\ INDEX$$

$$BF20 = 2.26 - EXCHANGE\ RATE$$

$$BF21 = 16.71 - ln\_EXPORTS$$

$$BF1BF4 = (UNEMPLOYMENT\ RATE - 7.3)(INTEREST\ RATE - 8.5392)$$

$$BF1BF5 = (UNEMPLOYMENT\ RATE - 7.3)(INDUSTRY\ PRODUCTION\ INDEX - 82.9933)$$

$$BF1BF7 = (UNEMPLOYMENT\ RATE - 7.3)(EXCHANGE\ RATE - 1.2)$$

$$BF1BF10 = (UNEMPLOYMENT\ RATE - 7.3)(\ln\_EXPORTS - 17.3359)$$

$$BF1BF12 = (UNEMPLOYMENT\ RATE - 7.3)(\ln\_IMPORTS - 17.14)$$

$$BF1BF17 = (UNEMPLOYMENT\ RATE - 7.3)(82.9933 - INDUSTRY\ PRODUCTION\ INDEX)$$

Table 7
ANOVA decomposition obtained by using AIC criterian.

Function	Std	AIC	BF	D.f.	Variables
1	0.011647	−363.1	1	2.5	Unemployment
2	0.011607	−363.5	1	2.5	Inflation
3	0.009101	−388.7	3	7.5	Industry production index
4	0.011466	−362.9	2	5.0	Exchange rate
5	0.011366	−366.0	1	2.5	Imports
6	0.011275	−367.0	1	2.5	Exports
7	0.009908	−382.5	1	2.5	Unemployment-Interest
8	0.007044	−419.4	3	7.5	Unemployment-Industry production index
9	0.008928	−393.0	2	5.0	Unemployment-Exchange rate
10	0.006905	−421.8	3	7.5	Unemployment-Exports
11	0.010707	−373.2	1	2.5	Unemployment-Imports

$$BF1BF18 = (UNEMPLOYMENT\ RATE - 7.3)(53.3 - INDUSTRY\ PRODUCTION\ INDEX)$$

$$BF1BF20 = (UNEMPLOYMENT\ RATE - 7.3)(2.26 - EXCHANGE\ RATE)$$

$$BF1BF21 = (UNEMPLOYMENT\ RATE - 7.3)(16.71 - \ln \_EXPORTS)$$

$$BF1BF22 = (UNEMPLOYMENT\ RATE - 7.3)(17.3359 - \ln \_EXPORTS)$$

MARS model obtained by usuing AIC criterian contains 19 BFs. As it can be seen from Table 7, the model includes 11 variables. The unemployment rate, inflation, industry production index, exchange rate, imports and exports are variables contributing to the model as main effect. It is seen that a model with fewer parameters is formed compared to the GCV criterion.

Results of ANOVA decomposition at significant variables depending on AIC criterion of MARS regression are obtained in Table 7. When we analyze the standard deviation values, we can say that unemployment and inflation variables have more effect on dependent variable. The fact that the AIC values of unemployment and exchange rate variables are larger than the values of other variables indicates how large the loss in the model is when these variables are removed from the model.

MARS model obtained with SBC criterian:

$$\begin{aligned} \mathrm{Y} = & -0.0002254 + 0.0005362\mathrm{BF3} - 0.000312\mathrm{BF5} - 0.000254\mathrm{BF6} + 0.0007473\mathrm{BF8} - 0.0005031\mathrm{BF11} + 0.0009622\mathrm{BF18} \\ & - 0.0000504\mathrm{BF20} - 0.0002505\mathrm{BF21} + 0.0001434\mathrm{BF1BF4} - 0.0005787\mathrm{BF1BF5} - 0.0002347\mathrm{BF1BF7} + 0.0003357\mathrm{BF1BF10} \\ & + 0.0002081\mathrm{BF1BF17} + 0.0000193\mathrm{BF1BF20} + 0.0001388\mathrm{BF1BF21} - 0.0000663\mathrm{BF1BF22} \end{aligned}$$

$$BF3 = INFLATION - 3.99$$

$$BF5 = INDUSTRY\ PRODUCTION\ INDEX - 82.9933$$

$$BF6 = INDUSTRY\ PRODUCTION\ INDEX - 53.3$$

$$BF8 = EXCHANGE\ RATE - 2.26$$

$$BF11 = ln\ \_IMPORTS - 17.8627$$

$$BF18 = 53.3 - INDUSTRY\ PRODUCTION\ INDEX$$

$$BF20 = 2.26 - EXCHANGE\ RATE$$

$$BF21 = 16.71 - ln\ \_EXPORTS$$

$$BF1BF4 = (UNEMPLOYMENT\ RATE - 7.3)(INTEREST\ RATE - 8.5392)$$

$$BF1BF5 = (UNEMPLOYMENT\ RATE - 7.3)(INDUSTRY\ PRODUCTION\ INDEX - 82.9933)$$

$$BF1BF7 = (UNEMPLOYMENT\ RATE - 7.3)(EXCHANGE\ RATE - 1.2)$$

$$BF1BF10 = (UNEMPLOYMENT\ RATE - 7.3)(\ln \_EXPORTS - 17.3359)$$

$$BF1BF17 = (UNEMPLOYMENT\ RATE - 7.3)(82.9933 - INDUSTRY\ PRODUCTION\ INDEX)$$

$$BF1BF20 = (UNEMPLOYMENT\ RATE - 7.3)(2.26 - EXCHANGE\ RATE)$$

Table 8
ANOVA decomposition obtained by using SBC criterion.

Function	Std	SBC	BF	D.f.	Variables
1	0.011607	−361.4	1	2.5	Inflation
2	0.009101	−382.4	3	7.5	Industry production index
3	0.011466	−358.8	2	5.0	Exchange rate
4	0.011366	−363.9	1	2.5	Imports
5	0.011275	−364.9	1	2.5	Exports
6	0.009908	−380.4	1	2.5	Unemployment-Interest
7	0.007779	−405.3	2	5.0	Unemployment-Industry production index
8	0.008928	−388.8	2	5.0	Unemployment-Exchange rate
9	0.006905	−415.5	3	7.5	Unemployment-Exports

$$BF1BF21 = (UNEMPLOYMENT\ RATE - 7.3)(16.71 - \ln\_EXPORTS)$$

$$BF1BF22 = (UNEMPLOYMENT\ RATE - 7.3)(17.3359 - \ln\_EXPORTS)$$

Penalty term of SBC based on Bayes theory is larger than the penalty term of AIC. Therefore, the model selected with SBC is expected to be smaller or equal in size than the model selected with AIC. The results of the analysis also support this. While a model containing 19 BFs is obtained using AIC, this value is obtained as 16 BFs by using SBC. And the model selected with SBC includes 9 variables (see Table 8). Inflation, industry production index, exchange rate, imports and exports are variables that contributed to the model (which is obtained with SBC) as main effect.

When the standard deviation values are examined, it is seen that the inflation and exchange rate variables included, respectively, in the model with 1 and 2 BFs have more effects on the dependent variable. The fact that the SBC values of the inflation and exchange rate variables are larger than the values of other variables can be interpreted as the loss in the model will be large when these variables are removed from the model.

MARS model obtained with $ICOMP(IFIM)_{PEU}$ criterion:

$$\begin{aligned} Y = {} & 0.0003294 - 0.0001451BF4 - 0.0001128BF5 + 0.0003227BF6 - 0.0000537BF18 + 0.0002261BF21 - 0.0002504BF1BF4 \\ & + 0.0001778BF1BF5 + 0.0002267BF1BF7 - 0.000201BF1BF17 - 0.0000099BF1BF21 \end{aligned}$$

$$BF4 = INTEREST\ RATE - 8.5392$$

$$BF5 = INDUSTRY\ PRODUCTION\ INDEX - 82.9933$$

$$BF6 = INDUSTRY\ PRODUCTION\ INDEX - 53.3$$

$$BF18 = 53.3 - INDUSTRY\ PRODUCTION\ INDEX$$

$$BF21 = 16.71 - ln\_EXPORTS$$

$$BF1BF4 = (UNEMPLOYMENT\ RATE - 7.3)(INTEREST\ RATE - 8.5392)$$

$$BF1BF5 = (UNEMPLOYMENT\ RATE - 7.3)(INDUSTRY\ PRODUCTION\ INDEX - 82.9933)$$

$$BF1BF7 = (UNEMPLOYMENT\ RATE - 7.3)(EXCHANGE\ RATE - 1.2)$$

$$BF1BF17 = (UNEMPLOYMENT\ RATE - 7.3)(82.9933 - INDUSTRY\ PRODUCTION\ INDEX)$$

$$BF1BF21 = (UNEMPLOYMENT\ RATE - 7.3)(16.71 - \ln\_EXPORTS)$$

The ICOMP(IFIM)$_{PEU}$ criterion, whose purpose is to express the most accurate model in the simplest way by providing the best balance between overfitting and lack-of-fit of the model, applies a more stricter penalty method for this purpose. Instead of penalizing

Table 9
ANOVA decomposition obtained by using ICOMP(IFIM)$_{PEU}$ criterion.

Function	**Std**	ICOMP(IFIM)$_{PEU}$	**BF**	**D.f.**	**Variables**
1	0.011331	−358.1	1	2.5	Interest
2	0.009101	−371.4	3	7.5	Industry production index
3	0.011275	−361.0	1	2.5	Exports
4	0.009908	−378.6	1	2.5	Unemployment-Interest
5	0.007626	−404.3	2	5.0	Unemployment-Industry production index
6	0.009290	−381.3	1	2.5	Unemployment-Exchange rate
7	0.010856	−368.4	1	2.5	Unemployment-Exports

Table 10
Comparison of MARS models performances for each criterion.

	MSE		R^2		
Criteria	Simulation	Real Data	Simulation	Real Data	avg number of BFs
GCV	1032	0,748[a]	0,952	0942[a]	17,5
AIC	0,704[a]	0,770	0968[a]	0,941	20
SBC	0,873	0881	0,960	0932	15,5
ICOMP(IFIM)$_{PEU}$	1256	0,756	0942	0,942[a]	9,5

[a] shows better performance.

the number of free parameters as in other criteria, it prefers to penalize the complexity of the covariance. Therefore, it is expected to create a model with less BF compared to other information criteria. The interest rate, industry production index and exports variables obtained using ICOMP(IFIM)$_{PEU}$ contribute as main effects. Moreover, when the interactive variables included in the model are examined, ICOMP(IFIM)$_{PEU}$ criterion determines a relationship between interest rates and unemployment and includes this relationship as a new variable in the model. We can say that ICOMP(IFIM)$_{PEU}$ is a successful criterion in model selection since the ICOMP(IFIM)$_{PEU}$ criterion contains 10 BFs. From Table 9 it can be seen that the model includes 7 variables.

We see that the interest rate and exports variables have more effect on the dependent variable. ICOMP(IFIM)$_{PEU}$ values of the interest rate and exports variables are larger than the values of other variables mean that the loss that will occur when these variables are removed from the model will be larger.

Finally, the mean squared error (MSE) and coefficient of determination, i.e. R-squared, (R^2) values are given to assess and compare the performances of MARS models for each criterion (see Table 10). Considering all the performance criteria, the model selected for the simulation dataset with the AIC criterion performs better than the other models. When the corresponding results for the real dataset are examined, it is seen that the model selected by GCV criterion is better than the other models. These results may be because of the MARS models selected by AIC and GCV criteria contain more number of BFs. In addition, for the real dataset, the coefficient of determination value of the model selected by ICOMP(IFIM)$_{PEU}$ criterion is equal to that of the model selected by GCV criterion, although ICOMP(IFIM)$_{PEU}$ criterion includes fewer BFs in the models. The R^2 value of the model selected by AIC criterion is very close to the R^2 values of the models selected by GCV and ICOMP(IFIM)$_{PEU}$ criteria. While the BFs number of the MARS model obtained by AIC criterion is the highest, the MARS model obtained by ICOMP(IFIM)$_{PEU}$ criterion is the simplest. It is better to choose a model with less BFs when there is not much difference between the measurement results.

6. Conclusion

MARS technique gives the relationship of independent variables and their interactions with the dependent variable. MARS technique uses GCV criterion for model selection. Choosing an arbitrary value for the smoothing parameter used in the formula of the GCV criterion and obtaining high-dimensional models using the GCV criterion are the criticized aspects of the MARS method. In this study, the model selection performance of GCV criterion is assessed and this criterion is compared with other information criteria using a simulation study and on a real NPL dataset. As a result of our analyzes, a MARS model is created, which includes 25 BFs with the GCV criterion, 19 BFs with the AIC criterion, 16 BFs with the SBC criterion, and 10 BFs with the ICOMP(IFIM)$_{PEU}$ criterion.

When the criteria are examined in terms of the variables chosen for the models;

- Unemployment rate, inflation, interest rate, industry production index, exchange rate, imports and exports are the main variables contributing to the model obtained with the GCV criterion.
- Unemployment, inflation, industry production index, exchange rate, imports and exports are the variables that contribute to the model obtained with the AIC criterion as the main effect.
- Inflation, industry production index, exchange rate, imports and exports are the variables that contribute to the model obtained with the SBC criterion as the main effect.
- Interest rate, industry production index and exports are the variables that contribute to the model obtained with the ICOMP(IFIM)$_{PEU}$ criterion as the main effect. Moreover, it has been observed that the unemployment variable interacts with the interest rate, industry production index and exchange rate variables and affects the NPL.

As a result, the outputs obtained from the study show that; the ICOMP(IFIM)$_{PEU}$ criterion, which enables the selection of the true model with less BF, can be used as a powerful and useful criterion in the MARS algorithm.

Author contribution statement

Meryem Bekar Adiguzel and Mehmet Ali Cengiz: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data.

Meryem Bekar Adiguzel: Performed the experiments; Analyzed and interpreted the data; wrote the paper.

Data availability statement

Data included in article/supplementary material/referenced in article.

Additional information

No additional information is available for this paper.

Funding

This research received no external funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

APPENDIX-1: Real data

Non Performing Loans	Unemployment Rate	Inflation	Interest Rate	Industry Production Index	Exchange Rate	Gross Domestic Product	Export	Import
5,82%	13,70	11,84	14,54	124,19	5,80	1.189.855.413,16	17.977.779,21	27.822.542,39
5,38%	13,80	9,26	20,64	113,02	5,68	1.145.099.354,34	17.927.740,87	28.427.445,19
4,71%	13,00	15,72	26,40	110,50	5,88	1.023.855.087,17	18.214.793,00	29.165.359,42
4,34%	14,10	19,71	24,18	105,39	5,37	922.029.155,99	19.019.136,60	30.200.269,07
4,15%	13,50	20,30	30,39	117,75	5,53	1.017.190.008,72	19.609.666,43	32.270.116,14
3,43%	11,40	24,52	28,66	112,17	5,59	1.026.648.922,65	21.013.530,91	35.640.203,97
3,20%	10,16	15,39	19,50	114,39	4,37	890.435.944,67	21.681.658,33	35.360.565,85
3,05%	10,12	10,23	17,49	111,72	3,82	790.113.059,53	23.189.347,54	36.353.567,28
3,10%	10,40	11,92	16,68	126,84	3,80	892.228.264,53	24.113.841,31	38.097.509,82
3,20%	10,60	11,20	16,58	112,12	3,52	833.706.740,87	25.975.232,65	40.647.773,12
3,25%	10,20	10,90	15,85	109,48	3,59	735.280.547,65	27.285.761,92	43.088.794,21
3,39%	11,70	11,29	14,71	102,05	3,70	649.434.601,86	30.038.030,95	47.418.207,72
3,41%	12,70	8,53	14,35	114,77	3,28	747.226.024,63	33.639.635,66	52.247.149,42
3,48%	11,30	7,28	14,78	95,29	2,97	666.176.429,47	34.994.753,71	55.387.122,89
3,46%	10,20	7,64	15,62	105,72	2,90	631.232.693,23	37.037.950,51	55.137.537,17
3,45%	10,10	7,46	16,00	97,85	2,95	563.890.602,00	26.285.719,14	39.078.009,06
3,25%	10,80	8,81	15,67	111,04	2,91	646.500.325,08	25.121.683,98	30.833.904,62
3,07%	10,30	7,95	14,59	97,90	2,85	631.512.354,71	22.984.527,13	33.015.698,07
3,01%	9,60	7,20	13,35	100,87	2,67	562.947.770,73	26.132.662,58	37.259.887,91
2,96%	10,60	7,61	12,77	90,19	2,46	497.687.043,16	27.512.471,84	38.918.668,13
2,97%	10,90	8,17	12,79	100,32	2,27	557.419.788,61	26.691.063,68	40.899.122,49
3,05%	10,50	8,86	12,19	94,09	2,17	548.625.834,71	28.444.820,35	43.726.839,94
2,86%	9,10	9,16	14,23	94,29	2,12	487.151.068,27	28.023.658,61	46.785.777,85
2,94%	9,70	8,39	14,79	88,26	2,22	451.269.184,24	30.770.060,02	54.261.741,39
2,87%	9,60	7,40	11,02	96,35	2,03	491.085.106,51	32.280.376,79	58.607.863,90
2,87%	9,20	7,88	10,97	89,20	1,97	491.263.793,92	33.694.262,21	61.259.818,69
2,89%	8,10	8,30	9,78	90,06	1,84	441.539.542,87	35.079.775,10	61.926.342,92
3,14%	9,40	7,29	11,58	80,39	1,79	385.824.643,40	33.797.330,25	57.785.923,55
3,04%	9,30	6,16	12,23	89,30	1,79	429.732.717,44	36.059.975,06	57.976.902,96
3,14%	8,30	9,19	14,79	82,99	1,80	424.705.390,32	38.537.663,95	59.772.547,36
2,83%	7,30	8,87	14,19	84,42	1,81	382.070.001,70	39.782.515,20	60.029.446,36
2,94%	9,10	10,43	14,74	77,38	1,80	333.164.005,46	37.950.624,72	57.542.787,84
2,92%	9,00	10,45	13,39	87,36	1,84	385.734.139,69	38.114.830,20	62.033.201,03
2,95%	8,20	6,15	11,48	80,56	1,73	381.898.595,37	37.678.061,58	65.085.408,69
3,20%	8,70	6,24	9,64	79,52	1,57	336.234.139,95	38.180.766,47	61.773.059,36
3,59%	10,10	3,99	8,54	72,76	1,58	290.610.290,52	38.252.883,95	63.568.884,77
4,08%	10,60	6,40	8,61	77,48	1,46	322.360.447,43	41.340.967,32	60.618.424,37
4,83%	10,60	9,24	8,91	70,14	1,52	318.732.806,05	39.251.904,34	61.082.515,71
5,10%	9,90	8,37	9,15	69,11	1,54	278.647.853,11	39.182.873,71	60.008.683,51
5,81%	12,80	9,56	8,96	60,81	1,51	240.272.871,66	37.493.874,08	60.807.202,91
6,40%	12,60	6,53	9,98	67,73	1,49	274.874.781,66	38.217.761,37	55.117.476,90
6,56%	12,50	5,27	12,19	63,04	1,50	271.840.902,50	35.685.145,70	53.049.625,45
6,04%	12,20	5,73	14,32	61,33	1,57	241.220.604,55	35.577.523,29	50.451.581,42
5,67%	14,70	7,89	19,87	53,29	1,66	211.255.559,34	34.150.724,66	48.381.281,87

(*continued on next page*)

(*continued*)

Non Performing Loans	Unemployment Rate	Inflation	Interest Rate	Industry Production Index	Exchange Rate	Gross Domestic Product	Export	Import
4,59%	12,70	10,06	21,67	62,55	1,54	258.624.782,99	35.200.766,54	48.450.877,07
3,74%	9,80	11,13	18,50	68,79	1,21	268.726.652,43	36.128.443,05	50.211.892,34
3,69%	8,70	10,61	17,65	72,38	1,26	249.483.189,18	35.048.448,89	49.867.810,66
3,69%	10,10	9,15	17,04	68,59	1,20	217.948.233,82	36.032.107,84	50.024.107,07
4,16%	9,90	8,39	17,64	71,22	1,19	240.942.451,97	38.481.147,60	51.649.351,64
4,27%	8,90	7,12	18,45	69,50	1,29	237.359.324,52	39.621.932,17	57.234.284,58
4,36%	8,30	8,60	19,19	69,46	1,34	212.565.694,63	39.078.155,26	61.458.594,06
4,41%	9,50	10,86	19,43	63,66	1,41	189.593.408,03	39.692.777,55	61.841.625,58
4,39%	9,50	9,65	20,22	67,49	1,46	217.970.763,48	41.534.955,88	63.833.440,74
4,44%	8,40	10,55	19,85	65,52	1,50	217.039.426,46	41.048.283,36	60.055.096,54
4,49%	8,10	10,12	16,84	65,51	1,46	193.081.372,66	43.007.491,14	52.506.712,88
5,33%	10,00	8,16	18,22	57,41	1,33	161.135.992,52	42.151.159,29	46.705.969,01
5,61%	10,40	7,72	18,54	64,89	1,36	185.766.662,29	43.030.367,14	50.924.472,61
6,21%	9,20	7,99	19,09	61,37	1,34	185.021.847,57	41.718.909,48	49.569.257,75
6,39%	8,70	8,95	20,74	58,90	1,36	161.871.869,54	43.424.691,56	50.534.605,13
6,85%	10,00	7,94	23,93	53,30	1,33	141.042.563,35	42.938.342,66	52.105.678,02

APPENDIX-2: Data with Ln Value

Non Performing Loans	Unemployment Rate	Inflation	Interest Rate	Industry Production Index	Exchange Rate	Gross Domestic Product (Ln)	Export (Ln)	Import (Ln)
5,82%	13,70	11,84	14,54	124,19	5,80	20,897	16,705	17,141
5,38%	13,80	9,26	20,64	113,02	5,68	20,859	16,702	17,163
4,71%	13,00	15,72	26,40	110,50	5,88	20,747	16,718	17,188
4,34%	14,10	19,71	24,18	105,39	5,37	20,642	16,761	17,223
4,15%	13,50	20,30	30,39	117,75	5,53	20,740	16,792	17,290
3,43%	11,40	24,52	28,66	112,17	5,59	20,750	16,861	17,389
3,20%	10,16	15,39	19,50	114,39	4,37	20,607	16,892	17,381
3,05%	10,12	10,23	17,49	111,72	3,82	20,488	16,959	17,409
3,10%	10,40	11,92	16,68	126,84	3,80	20,609	16,998	17,456
3,20%	10,60	11,20	16,58	112,12	3,52	20,541	17,073	17,520
3,25%	10,20	10,90	15,85	109,48	3,59	20,416	17,122	17,579
3,39%	11,70	11,29	14,71	102,05	3,70	20,292	17,218	17,675
3,41%	12,70	8,53	14,35	114,77	3,28	20,432	17,331	17,771
3,48%	11,30	7,28	14,78	95,29	2,97	20,317	17,371	17,830
3,46%	10,20	7,64	15,62	105,72	2,90	20,263	17,427	17,825
3,45%	10,10	7,46	16,00	97,85	2,95	20,150	17,085	17,481
3,25%	10,80	8,81	15,67	111,04	2,91	20,287	17,039	17,244
3,07%	10,30	7,95	14,59	97,90	2,85	20,264	16,950	17,312
3,01%	9,60	7,20	13,35	100,87	2,67	20,149	17,079	17,433
2,96%	10,60	7,61	12,77	90,19	2,46	20,025	17,130	17,477
2,97%	10,90	8,17	12,79	100,32	2,27	20,139	17,100	17,527
3,05%	10,50	8,86	12,19	94,09	2,17	20,123	17,163	17,593
2,86%	9,10	9,16	14,23	94,29	2,12	20,004	17,149	17,661
2,94%	9,70	8,39	14,79	88,26	2,22	19,928	17,242	17,809
2,87%	9,60	7,40	11,02	96,35	2,03	20,012	17,290	17,886
2,87%	9,20	7,88	10,97	89,20	1,97	20,012	17,333	17,931
2,89%	8,10	8,30	9,78	90,06	1,84	19,906	17,373	17,941
3,14%	9,40	7,29	11,58	80,39	1,79	19,771	17,336	17,872
3,04%	9,30	6,16	12,23	89,30	1,79	19,879	17,401	17,876
3,14%	8,30	9,19	14,79	82,99	1,80	19,867	17,467	17,906
2,83%	7,30	8,87	14,19	84,42	1,81	19,761	17,499	17,910
2,94%	9,10	10,43	14,74	77,38	1,80	19,624	17,452	17,868
2,92%	9,00	10,45	13,39	87,36	1,84	19,771	17,456	17,943
2,95%	8,20	6,15	11,48	80,56	1,73	19,761	17,445	17,991
3,20%	8,70	6,24	9,64	79,52	1,57	19,633	17,458	17,939
3,59%	10,10	3,99	8,54	72,76	1,58	19,487	17,460	17,968
4,08%	10,60	6,40	8,61	77,48	1,46	19,591	17,537	17,920
4,83%	10,60	9,24	8,91	70,14	1,52	19,580	17,486	17,928
5,10%	9,90	8,37	9,15	69,11	1,54	19,445	17,484	17,910
5,81%	12,80	9,56	8,96	60,81	1,51	19,297	17,440	17,923
6,40%	12,60	6,53	9,98	67,73	1,49	19,432	17,459	17,825
6,56%	12,50	5,27	12,19	63,04	1,50	19,421	17,390	17,787
6,04%	12,20	5,73	14,32	61,33	1,57	19,301	17,387	17,737

(*continued on next page*)

(*continued*)

Non Performing Loans	Unemployment Rate	Inflation	Interest Rate	Industry Production Index	Exchange Rate	Gross Domestic Product (Ln)	Export (Ln)	Import (Ln)
5,67%	14,70	7,89	19,87	53,29	1,66	19,169	17,346	17,695
4,59%	12,70	10,06	21,67	62,55	1,54	19,371	17,377	17,696
3,74%	9,80	11,13	18,50	68,79	1,21	19,409	17,403	17,732
3,69%	8,70	10,61	17,65	72,38	1,26	19,335	17,372	17,725
3,69%	10,10	9,15	17,04	68,59	1,20	19,200	17,400	17,728
4,16%	9,90	8,39	17,64	71,22	1,19	19,300	17,466	17,760
4,27%	8,90	7,12	18,45	69,50	1,29	19,285	17,495	17,863
4,36%	8,30	8,60	19,19	69,46	1,34	19,175	17,481	17,934
4,41%	9,50	10,86	19,43	63,66	1,41	19,060	17,497	17,940
4,39%	9,50	9,65	20,22	67,49	1,46	19,200	17,542	17,972
4,44%	8,40	10,55	19,85	65,52	1,50	19,196	17,530	17,911
4,49%	8,10	10,12	16,84	65,51	1,46	19,079	17,577	17,776
5,33%	10,00	8,16	18,22	57,41	1,33	18,898	17,557	17,659
5,61%	10,40	7,72	18,54	64,89	1,36	19,040	17,577	17,746
6,21%	9,20	7,99	19,09	61,37	1,34	19,036	17,546	17,719
6,39%	8,70	8,95	20,74	58,90	1,36	18,902	17,587	17,738
6,85%	10,00	7,94	23,93	53,30	1,33	18,765	17,575	17,769

References

[1] J.H. Friedman, Multivariate adaptive regression splines, Ann. Stat. 19 (1) (1991) 1–67.

[2] E.K. Koç, H. Bozdogan, Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function, Mach. Learn. 101 (1–3) (2014) 35–58.

[3] N.B. Serrano, A.S. Sanchez, F.S. Lasheras, F.J.I. Rodriguez, G.F. Valverde, Identification of Gender Differences in the Factors Influencing Shoulders, Neck and Upper Limb MSD by Means of Multivariate Adaptive Regression Splines (MARS), Applied Ergonomics, 2020, https://doi.org/10.1016/j.apergo.2019.102981.

[4] S. Nacar, M. San, M. Kankal, U. Okkan, Climate change projections of temperature and precipitation in the Eastern Black Sea Basin, Turkey by using multivariate adaptive regression splines statistical downscaling method, Research Square (2021), https://doi.org/10.21203/rs.3.rs-647619/v1.

[5] N. Charron, P. Annoni, What is the Influence of news media on people's perception of corruption? Parametric and Non Parametric approaches. Springer Link, Soc. Indicat. Res. 153 (2021) 1139–1165.

[6] D. Guo, H. Chen, L. Tang, Z. Chen, P. Samui, Assessment of rockburst risk using multivariate adaptive regression splines and deep forest model, Prog. Earth Planet. Sci. 17 (2022) 1183–1205.

[7] F. Saadaoui, M. Khalfi, Revisiting Islamic banking efficiency using multivariate adaptive regression splines, Ann. Oper. Res. (2022), https://doi.org/10.1007/s10479-022-04545-2.

[8] D. Sabancı, M.A. Cengiz, Random ensemble MARS: model selection in multivariate adaptive regression spline using random forest approach, Journal of New Theory 40 (2022) 27–45, https://doi.org/10.53570/jnt.1147323.

[9] E. Pamukcu, Choosing the optimal hybrid covariance estimators in adaptive elastic net regression models using information complexity, J. Stat. Comput. Simulat. 89 (2019) 2983–2996.

[10] Y. Susanti, H. Pratiwi, S. Sulistijowati, T. Liana, M estimation, S estimation and MM estimation in robust regression, Int. J. Pure Appl. Math. 91 (3) (2014) 349–360.

[11] Y. Güney, H. Bozdogan, O. Arslan, Robust model selection in linear regression models using information complexity, J. Comput. Appl. Math. 398 (2021), https://doi.org/10.1016/j.cam.2021.113679.

[12] E. Russel, Wamiliana, Nairobi, Warsono, M. Usman, J.I. Daoud, Dynamic Modeling and Forecasting Data Energy Used and Carbon Dioxide (CO_2), Science and Technology Indonesia, 2022, https://doi.org/10.26554/sti.2022./.2.228-23/.

[13] P.B. Gohain, M. Jansson, Robust information criterion for model selection in sparse high-dimensional linear regression models, Electrical Engineering and Systems Science (2022), https://doi.org/10.48550/arXiv.2206.08731.

[14] A. Ozmen, *Robust conic Quadratic Programming Applied to Quality Improvement - a Robustification of CMARS*, Master Thesis, Middle East Technical University, Department of Scientific Computing, Ankara, 2010, p. 139.

[15] T. Hastie, R. Tibshirani, J.H. Friedman, The Element of Statistical Learning, Springer Verlag, New York, 2001.

[16] J. Stevens, An Investigation of Multivariate Adaptive Regression Splines for Modeling and Analysis of Univariate and Semi-multivariate Time Series Systems, Ph. D. Thesis, Naval Postgraduate School, 1991.

[17] M. Kriner, *Survival Analysis with Multivariate Adaptive Regression Splines*, Dissertation, LMU München, Faculty of Mathematics, Computer Science and Statistics, 2007.

[18] P. Craven, G. Wahba, Smoothing noisy data with spline functions, Numerische Mathematik, Verlag 31 (1979) 377–403.

[19] A.R. Barron, X. Xiao, Discussion: multivariate adaptive regression splines, Ann. Stat. 19 (1991) 67–82.

[20] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle. Second International Symposium on Information Theory, Academiai Kiado, Budapest, 1973, pp. 267–281.

[21] S.L. Sclove, Application of model selection criteria to some problems in multivariate analysis, Journal of the Psychometric Society 52 (1987) 333–343.

[22] G. Schwartz, Estimating the dimension of model, Ann. Stat. 6 (1978) 461–464.

[23] H. Bozdoğan, ICOMP: A New Model Selection Criteria. *Classification and Related Methods of Data Analysis*, 1988, pp. 599–608.

[24] H. Bozdoğan, On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models, Commun. Stat. Theor. Methods 19 (1990) 221–278.

[25] H. Bozdoğan, D.M.A. Haughton, Informational complexity criteria for regression models, Comput. Stat. Data Anal. 28 (1998) 51–76.

[26] H. Bozdoğan, Akaike's information criterion and recent developments in information complexity, J. Math. Psychol. 28 (2000) 51–76.

[27] H. Bozdoğan, A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation, Istanbul University Journal of the School and Business Administration 39 (2010) 370–398.

[28] Rstudio. http://www.rstudio.com.

[29] MARS, Salford Systems. http://www.salfordsystems.com/mars/phb).

[30] MARS User Guide, Salford Systems, San Diego, CA, 2001.