J Ind Eng Int
DOI 10.1007/s40092-017-0238-2

ORIGINAL RESEARCH

Modeling and forecasting US presidential election using learning algorithms

Mohammad Zolghadr[1] · Seyed Armin Akhavan Niaki[2] · S. T. A. Niaki[1]

Received: 20 June 2016 / Accepted: 15 September 2017

Abstract The primary objective of this research is to obtain an accurate forecasting model for the US presidential election. To identify a reliable model, artificial neural networks (ANN) and support vector regression (SVR) models are compared based on some specified performance measures. Moreover, six independent variables such as GDP, unemployment rate, the president's approval rate, and others are considered in a stepwise regression to identify significant variables. The president's approval rate is identified as the most significant variable, based on which eight other variables are identified and considered in the model development. Preprocessing methods are applied to prepare the data for the learning algorithms. The proposed procedure significantly increases the accuracy of the model by 50%. The learning algorithms (ANN and SVR) proved to be superior to linear regression based on each method's calculated performance measures. The SVR model is identified as the most accurate model among the other models as this model successfully predicted the outcome of the election in the last three elections (2004, 2008, and 2012). The proposed approach significantly increases the accuracy of the forecast.

✉ S. T. A. Niaki
Niaki@Sharif.edu

Mohammad Zolghadr
mohammad.zolghadr.70@gmail.com

Seyed Armin Akhavan Niaki
arniaki@mix.wvu.edu

[1] Department of Industrial Engineering, Sharif University of Technology, P.O. Box 11155-9441, Azadi Ave, Tehran, Iran

[2] Department of Statistics, Eberly College of Arts and Sciences, West Virginia University, Morgantown, USA

Keywords Presidential election · Forecasting · Artificial neural network · Support vector regression · Linear regression

Introduction

The United States presidential election is among influential factors on not only the local market but also the global economy. Researchers must pay more attention to political events and how they influence the development of competitive local markets alongside their influence on the global economy. Given the significance of the US Presidential Election and how it is capable of influencing the global economy, a bulk of scholars and politicians in the US have attempted to predict the outcome of the elections to formulate policies based on the obtained forecasts.

Modeling a complex phenomenon such as an election is neither a simple nor easy task. In some elections, the mechanism of the election is complicated, and in others, the candidates present further complexities in modeling the event. However, the US presidential election presents a slightly less difficult challenge. The bipartisanship of the political system in the United States presents a simple situation in which the failure of the incumbent party can be considered as the success of another party. Most forecasts have chosen the incumbent votes as the dependent variable in their models primarily due to this reason. This choice is based on the theory that the US presidential election is a referendum on the policies of the incumbent party. This theory states that people who are satisfied with the incumbent party are inclined to vote for their party's candidate, and people who are not satisfied are enthusiastic to vote for the opposing party's candidate.

Published online: 22 September 2017

The primary objective of this paper is to model and forecast the United States presidential election via the usage of learning algorithms. Political and economic variables are utilized in the model, and significant variables are identified through further analysis and statistical procedures. The dependent variable is defined as the electoral votes of the incumbent party. The incumbent party is considered as the dependent variable, because it presents further related variables such the incumbent president's approval rate and gross domestic product (GDP).

Increasing the accuracy of the obtained forecasts is another research objective. Moreover, analytical parsimonious models are desired to provide forecasts and different utilized learning algorithms are further compared based on some specified performance measures. The differences between artificial neural networks (ANN) and support vector regression (SVR) are investigated further based on two measures of error: mean absolute prediction error (MAPE), and root-mean-squared error (RMSE). Investigating the impacts of data mining techniques in increasing forecasting accuracy is another objective of this research, where four sets of data are examined using each technique.

The organization of this paper is as follows. The literature on US presidential election forecasts is thoroughly investigated and examined in the next section. In the third section, brief backgrounds on ANN and SVR are presented. The fourth section demonstrates the modeling process and the obtained forecasts using the above-mentioned algorithms alongside some utilized data mining techniques. The results of the best model of each method are compared, and furthermore, inferences about the effects of utilizing data mining techniques and learning algorithms are made in this section. Finally, we conclude the paper in "Results" section, where some recommendations and possibilities for future studies are presented.

Literature review

Although forecasting has been used many times in numerous fields, it has a brief history in political science. Forecasting political events started in the late 1970s when Fair (1978) investigated the effect of the economic condition in the election year as well the incumbent parties in a forecasting model. Sigelman (1979) examined the relation between the results of an election with the previous ones. Lewis-Beck and Rice (1982) developed a model using the president's job approval and an economic factor as independent variables. Abramowitz (1988) added a time-dependent variable to improve the performance of the forecasting model. The dependent variable of the model was the percentage of the incumbent party votes, and the independent variables were GDP growth, the incumbent president's job approval rating in June of the election year, and the consecutive terms that the incumbent party governs the country. He used the ordinary least-squares (OLS) method to estimate the parameters of the linear regression model. Later, Abramowitz (2016) utilized his model, which is called "Time for change forecasting model", to forecast 2016 election.

Some years later, Lewis-Beck and Rice (1992) reformed their model by adding two new variables: the result of the previous congress election and the previous presidential election. Holbrook and DeSart (1999) used the percentage of voters and the last votes of parties, as variables in their forecasting model. They employed the OLS method to estimate the parameters of their regression model.

Wlezien and Erikson (2004) introduced a model using economic indices and the percentage of the incumbent party votes as variables. They used the R^2 (coefficient of multiple determination) and the adjusted R^2 to evaluate the accuracy of their forecasting model. Later, Erikson and Wlezien (2016) employed their model to forecast 2016 presidential election by adding the polls to their model. An important research on forecasting the United States presidential election was conducted by Berg and Rietz (2014). These individuals who were political science professors in the University of Iowa proposed a method to predict the presidential election which is known as the Iowa prediction market. Lewis-Beck and Tien (2014) addressed the issue of forecasting from statistical models, and the way they might be improved. They used a real-world example on the US presidential elections in their work. They provided a summary of various leading US presidential election models that use various independent variables such as presidential popularity, GNP growth, primary support, house party advantage, peace and prosperity, and incumbency.

Fair (2011) allocated a chapter of his book to predicting the result of the US presidential election. The variables in his model were GNP, inflation rate, and the consecutive terms that the incumbent party governs the country and the percentage of the incumbent party votes. De Neve (2014) used data from the 1920 presidential election to the 2008 presidential election to forecast the result of the US presidential elections. The independent variables in his model were personal income growth rate, taxes, GNP, inflation rate, and unemployment rate. Interested readers are referred to Lewis-Beck (2005) on the principles and the practices of election forecasting.

Serious efforts have been undertaken to develop election forecasting in other countries. Ford et al. (2016) developed a three-stage method to forecast parliamentary election results from vote preferences in British opinion polls. Rallings et al. (2016) introduced a model using local government election results to estimate a national

equivalent vote in the UK parliamentary election. An important research on forecasting the 2013 German Bundestag Election was conducted by Munzert (2017). He used a time-series method to forecast 2013 German election based on many polls and historical election results. Charles and Reid (2016) also used election results and macroeconomic variables from 1962 to 2015 to develop a time-series model to forecast the 2016 General Election in Jamaica.

Learning algorithms

This section provides brief backgrounds on two learning algorithms, i.e., support vector regression and artificial neural network, utilized in this paper to forecast US presidential election.

Support vector regression

The support vector (SV) algorithm is a nonlinear generalization of the generalized portrait algorithm proposed for the first time by Vapnik and Lerner (1963) and Vapnik and Chervonenkis (1964) in the 1960s. It is based on the theory of statistical learning that has been developed over the last 3 decades by Vapnik and Chervonenkis (1974) and Vapnik (1982, 1995). The statistical learning theory, in essence, characterizes properties of learning machines to make them able of generalization to unseen data. The SV has many applications including regression and time-series predictions and its excellent performance has been shown in Műller et al. (1997), Drucker et al. (1997), Stitson et al. (1999), and Mattera and Haykin (1999).

Suppose that the training data $\{(x^1, y^1), \ldots\ldots\ldots\ldots, (x^l, y^l)\} \subset X \times \mathbb{R}$, in which X is the space of the input parameters, are available. One possible realization of the training data set is the exchange rates of a currency measured in subsequent days along with their corresponding econometric indicators. The goal in ε-SV regression is to find a function $f(x)$ with the most ε-deviation from the obtained targets y^l for all the training data, and at the same time as flat as possible (Vapnik 1995).

Depending on the form of the function $f(x)$, support vector regression (SVR) is classified into two classes of linear and nonlinear SVR that are discussed as follows.

Linear SVR

In linear SVR, the function $f(x)$ takes the form:

$$f(x) = <w, x> + b, \tag{1}$$

where $w \in X$ is the slope, $b \in \mathbb{R}$ is the y-intercept, and $<.,.>$ denotes the dot product in X. Moreover, the flatness, in this case, means that small w is desired. An alternative to having small w is to minimize the norm $\|w^2\| = <w, w>$. In other words, the following convex optimization problem is involved:

$$\begin{aligned} &\text{Min}\, \frac{1}{2}\|w\|^2 \\ &s.t. \begin{cases} y^l - <w,\ x^l> - b \le \varepsilon \\ -y^l + <w,\ x^l> + b \le \varepsilon. \end{cases} \end{aligned} \tag{2}$$

The implicit assumption in (2) is that there is a function $f(x)$, such that the above convex optimization problem is feasible. However, sometimes, this may not be the case, for which one can introduce some slack variables $\xi_i^- \ge 0$ and $\xi_i^+ \ge 0$ to cope with infeasible constraints (Cortes and Vapnik 1995). This leads to the formulation stated in Vapnik (1995) as follows:

$$\begin{aligned} &\text{Min}\, \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i^- + \xi_i^+\right) \\ &s.t. \begin{cases} y^l - <w,\ x^l> - b \le \varepsilon + \xi_i^- \\ -y^l + <w,\ x^l> + b \le \varepsilon + \xi_i^+, \end{cases} \end{aligned} \tag{3}$$

where the constant $C > 0$ determines the trade-off between the flatness of the regression function $f(x)$ and the threshold up to which, deviations larger than ε are tolerated. In fact, $\frac{1}{2}\|w^2\|$ demonstrates the complexity of the model, and $C\sum_{i=1}^{l}\left(\xi_i^- + \xi_i^+\right)$ is defined as the empirical error of the model.

To solve the optimization problem stated in (3), a Lagrangian function is constructed for the objective function (the primal objective function) and the corresponding constraints as follows:

$$\begin{aligned} L = &\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i^- + \xi_i^+\right) \\ &- \sum_{i=1}^{l}\left(\lambda_i^+\xi_i^+ + \lambda_i^-\xi_i^-\right) \\ &- \sum_{i=1}^{l}\alpha_i^-\left(\varepsilon + \xi_i^- - y^i + <w, x^i> + b\right) \\ &- \sum_{i=1}^{l}\alpha_i^+\left(\varepsilon + \xi_i^+ + y^i - <w, x^i> - b\right), \end{aligned} \tag{4}$$

where λ_i^+, λ_i^-, α_i^-, and α_i^+ are Lagrangian multipliers. Then, using the saddle point condition obtained by the partial derivatives of the Lagrangian function with respect to the primal variables w, b, ξ_i^-, and ξ_i^+ and taking the advantage of the dual equivalence of the optimization problem at hand, the following optimization problem is solved easier:

$$\begin{aligned}
&\text{Max } w(\alpha_i^-, \alpha_i^+) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i^- + \xi_i^+) \\
&\quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^- - \alpha_i^+)\left(\alpha_j^- - \alpha_j^+\right) < x_i, x_j > \\
&\quad +\sum_{i=1}^{l}(\alpha_i^-(y^i - \varepsilon) - \alpha_i^+(y^i + \varepsilon)) = \\
&s.t.\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i^- + \xi_i^+) \\
&\quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^- - \alpha_i^+)\left(\alpha_j^- - \alpha_j^+\right) < x_i, x_j > \\
&\quad +\sum_{i=1}^{l}(\alpha_i^-(y^i - \varepsilon) - \alpha_i^+(y^i + \varepsilon)) \\
&s.t.\ 0 \le \alpha_i^-, \alpha_i^+ \le C \\
&\quad \sum_{i=1}^{l}(\alpha_i^- - \alpha_i^+) = 0.
\end{aligned} \tag{5}$$

Finally, by exploiting the Krauch–Kuhn–Tucker condition (Karush 1939; Kuhn and Tucker 1951), the solution of the dual problem is obtained as follows:

$$b = -\frac{1}{2} < \overline{w}, (x_r + x_s) > \tag{6}$$

$$w = \sum_{i=1}^{l}(\alpha_i^- - \alpha_i^+)x. \tag{7}$$

Nonlinear SVR

As seen in "Linear SVR" section, the goal in ε-SV regression is to find a function with the most ε-deviation from the obtained targets for all the training data, and at the same time as flat as possible. Sometimes, however, the linear regression function used in the linear SVR is not appropriate. For instance, when the inputs present nonlinear characteristics, their linearization reduces the accuracy of the model. A better choice is to use the Kernel function in the case where nonlinear characteristics are detected in the input values. Kernel functions preprocess the inputs, thus taking the nonlinear patterns into consideration during the preprocessing procedures.

Kernel functions are responsible for mapping inputs onto a feature space. Consider a nonempty set X. Then, the mapping

$$k : X \times X \to K \tag{8}$$

is a Kernel on X, if space H exists, in which the K-Hilbert and the map are as follows:

$$\begin{aligned}
&\varphi : X \to H \quad \text{and} \\
&\forall x, x' \in H, K(x, x') = < \varphi(x'), \varphi(x) > ,
\end{aligned} \tag{9}$$

where φ is the feature map and H is the feature space of K.

Using the Kernel function to map the feature space, the regression function is restated as follows:

$$y(x) = \sum_{i=1}^{l}(\alpha_i^- - \alpha_i^+).K(x^i, x) + b. \tag{10}$$

While several Kernel functions are available, the following are the wide known (Murphy 2012).

- Linear Kernel Function:

$$K(x, x^i) = < x, x^i > .$$

- Hyperbolic Tangent (Sigmoid) Kernel function:

$$K(x, x^i) = \tanh(\beta + \Upsilon < x^i, x >).$$

- Radial Basis Kernel function:

$$K(x, x^i) = \exp(-\Upsilon \|x - x^i\|^2).$$

To utilize linear regressions for nonlinear models, nonlinear maps that transform data into a multi-dimensional feature space are engaged. Thus, taking advantage of the dual problem, the following optimization problem is solved to find the optimal solution of the nonlinear SV problem:

$$\begin{aligned}
&\text{Max}_{\alpha^-, \alpha^+} w(\alpha^-, \alpha^+) = \text{Max}_{\alpha^-, \alpha^+} \sum_{i=1}^{l} \alpha_i^+(y^i - \varepsilon) - \alpha_i^-(y^i + \varepsilon) \\
&-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)K(x^i, x^j) \\
&s.t.\ 0 \le \alpha_i^-, \alpha_i^+ \le C \quad \sum_{i=1}^{l}(\alpha_i^- - \alpha_i^+) = 0.
\end{aligned} \tag{11}$$

The regression function is further defined by solving the prior problem using the Lagrangian method as follows:

$$f(x) = \sum_{SVs}(\overline{\alpha_i^-} - \overline{\alpha_i^+})K(x^i, x) + \overline{b}, \tag{12}$$

where

$$\overline{b} = -\frac{1}{2}\sum_{i=1}^{l}(\overline{\alpha_i^-} - \overline{\alpha_i^+})(K(x^i, x^r) + K(x^i, x^s)) \quad (13)$$

and

$$< \overline{w}, x > = \sum_{i=1}^{l}(\overline{\alpha_i^-} - \overline{\alpha_i^+})K(x^i, x^j). \quad (14)$$

Interested readers are referred to Murphy (2012) for more details.

Neural networks

Artificial neural network (ANN) is a relatively newly developed tool that has been widely employed for forecasting in various fields. An artificial neural network (ANN) is a system consisted of numerous simple parts that are in relation with each other. Data are processed using dynamic answers to the independent inputs in such networks. Applications have been increased after neural networks were able to solve indissoluble problems in recent years. For instance, Yousefi et al. (2015) used ANN to model the nonlinearity of wind speed to accurately forecast wind speed in wind farms. Markopoulos et al. (2016) compared the performances of various ANNs in predicting surface roughness. Maleki et al. (2015) employed an ANN to provide a step-change point estimation of the multi-attribute process variability. Shokrollahpour and Hosseinzadeh Lotfi (2016) integrated an ANN with DEA to determine the relative efficiency of one of the Iranian commercial bank branches. Bashiri et al. (2013) proposed an ANN approach to optimize uncorrelated multi-response problems with "smaller the better" type controllable factors. A comprehensive review on using ANNs as a forecasting tool was provided by Zhang et al. (1998).

There are four main points that justify the use of ANNs to forecast presidential elections; (1) ANNs are nonlinear, i.e., they can capture nonlinear relations between independent (input or feature) and dependent (output or response) variables, (2) ANNs are data driven, i.e., no explicit assumption on the model between the inputs and outputs is needed, (3) ANNs are able to generalize, i.e., they can produce good results even when they face to new input patterns, and (4) unlike statistical techniques, ANNs do not need assumptions on the distribution of input data (Niaki and Hoseinzade 2013). However, before their use, one must pay attention that sometimes, the robustness of their outcomes is questionable (Saad et al. 1998). Besides, they have three main disadvantages; (1) the determination of the optimal combination of the network parameters such as learning rate, momentum, number of hidden layers, number of hidden nodes in each layer, etc., is difficult, (2) selecting the relevant features of an ANN is not an easy job, and (3) great volume of data is required to train the network to achieve an accurate result (Zhu et al. 2008).

The network topology, the number of layers, the number of nodes in each layer, the activation function, and the learning algorithms are to be determined to design an appropriate ANN for a particular problem. Based on the topology, ANNs are mainly classified into two groups of feed-forward and recurrent networks. As the use of the recurrent topology in more common in univariate forecasting analysis (Saad et al. 1998), it will be used in this paper to forecast US presidential election.

Depending on the complexity of the problem, the number of network layers varies. Besides, many recurrent networks have one or more hidden layers in addition to the input and the output layers that are essential for an ANN design. As the available methods to determine the optimal number of hidden layers and hidden nodes are very complex and hard to apply (Zhang et al. 1998), in this paper, the common practice of identifying the proper network design, which is comparing the performances of ANNs with different designs and selecting the network that results in the best performance, is taken (Hosseini et al. 2006).

The input layer of an ANN consists of the input variables (features) that seem to be influential to the output variable. In this paper, these influential features are determined using the regression analysis, where the features of the proposed ANN are the potential independent variables.

The output layer of an ANN consists of nodes associated with the dependent variables. As the objective of this research is to forecast the outcome of US presidential election, the output layers of the proposed ANN consist of only one node.

The tangent hyperbolic sigmoid (Tansig) function as the most common one in the relevant literature is used as the activation function for the nodes of all layers. Furthermore, the error back-propagation algorithm is employed to train the designed ANN.

To design, train, and simulate the proposed ANN, the neural network toolbox of the MATLAB 7 package software is used in this research. Interested readers are referred to Demuth and Beale (1998) for a detailed description of this neural network toolbox.

Model development

To initialize the model development, preprocessing methods are performed. The utilized preprocessing methods in this research are (1) data transformation, (2) data reduction, and (3) clustering. Furthermore, SVR and ANN are the employed learning algorithms for the obtained forecasting

models. The acquired results of the prior mentioned algorithms are further compared to linear regression results based on the following measures of performance:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - P_i)^2} \tag{15}$$

$$\mathrm{MAPE} = 100 \times \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - P_i|}{Y_i}, \tag{16}$$

where Y_i is the observed result and P_i is the predicted result.

The dependent variable in this research is observed as the electoral votes of the incumbent party in 16 data sets. The forecasting model is developed based on the US presidential election data from 1952 to 2012, where the last three data sets out of the 16 have been set aside to validate the model. Furthermore, the potential independent variables are considered as follows:

- The number of the consecutive terms the incumbent party has been in office.
- Personal income.
- Electoral votes of the incumbent party in the previous election.
- Votes of the incumbent party in the last senate election.
- Votes of the incumbent party in the last house of representatives election.
- The president's approval rate.
- Unemployment rate.
- The number of times that the 3-month GDP is above 3.2 within the last 4 years.

For the data reduction process, the stepwise regression is performed, based on which the most significant variables are identified and selected for the model. The SPSS software is utilized to obtain the results in Table 1 using the stepwise method. The results indicate that the president's job approval rate is the only significant variable among the above-mentioned variables. The calculated adjusted *R*-square of the model is 0.714, which indicates that the model is only able to account for 0.714 of the variation of the dependent variable. Subsequently, the independent variables are altered, to obtain a better performing model with higher adjusted *R*-square values (Adj. *R*-square > 0.8).

Since the president's job approval has been identified as the only significant variable, the model is further reformed based on this finding. In the previous model, the president's job approval rate at the end of June of the election year was considered. However, this rate at the end of each month presents more data points for this significant variable. The president's job approval rate at the end of the first 8 months of the election year is thus utilized. Using the stepwise regression once again, the results are obtained in Table 2 by employing SPSS.

It is evident that the president's job approval at the end of April (VAR4) and June (VAR6) has been selected by the stepwise regression method. The calculated adjusted *R*-square value is 0.774 in this case which suggests that further improvements are necessary to obtain an acceptable model.

The next applied preprocessing method is data transformation. Data transformation is necessary for learning algorithms, since it prevents the algorithm from accentuating the variables with bigger data. In addition, it significantly reduces the error of the model. The most useful method in data transformation is the Mini–Max method. In this method, an interval for the data is taken into consideration. Considering (0, 1) or (−1, 1) is common practice; however, (0.3, 0.5) is specified as the interval for the Mini–Max transformations in this research based on a pilot study. Minifying intervals in data transformation leads to an extreme reduction in the error of the model. The main objective of learning algorithms is to find the optimal plane in the feasible space of the problem. Furthermore, the error is reduced due to the relative ease of searching a smaller and more limited space compared to the initial space. Another advantage of minifying intervals is related to the sigmoid functions in ANNs. Using this function in neural networks is also a common practice, and the derivative of a sigmoid function is used in the learning process. Since the derivative of this function near 0 and 1 is about 0, enlarging intervals might lead to divergence in the neural network algorithm. By utilizing the Mini–Max method, considering (0.3, 0.5) as the interval prior to the stepwise regression for the model, the obtained results in Table 3 suggest an improvement in the adjusted *R*-square value of the model to 0.782.

Clustering is another utilized data transformation method. Clustering is useful in decreasing noise in data, and it also increases the focus throughout the data within different clusters. To apply this method, the K-means algorithm is utilized. In this algorithm, K clusters are specified where the goal is to minimize this number to avoid possible resulting divergent systems. Table 4 demonstrates the results of this transformation method.

Table 1 Result of the first stepwise regression

Model	Variables entered	*R*-square	Adjusted *R*-square	Std. error of the estimate
1	President job approval 2	0.733	0.714	80.37846

Table 2 Result of the second stepwise regression

Model	Variables entered	*R*-square	Adjusted *R*-square	Std. error of the estimate
2	VAR4, VAR6	0.804	0.774	71.39166

Table 3 Model summary after applying data transformation

Model	*R*-square	Adjusted *R*-square	Std. error of the estimate
3	0.811	0.782	0.02945

Table 4 Clustering results

Case number	Cluster	Distance
1	1	0.074
2	2	0.073
3	3	0.021
4	3	0.021
5	2	0.067
6	1	0.098
7	1	0.097
8	1	0.118
9	1	0.064
10	1	0.038
11	2	0.050
12	1	0.075
13	1	0.030
14	1	0.045
15	2	0.048
16	1	0.021

As seen in the above table, the validation data (14th, 15th, and 16th data sets) are in the first and second clusters. Thus, it is concluded that the data in the third cluster are not useful, and thus, it is omitted. Using the prior preprocessing methods, the following four types of data sets are further used for the application of SVR and ANN learning algorithms in the next two subsections:

1. Data set 1: The initially transformed data (16 data sets, 8 variables).
2. Data set 2: Reduction and transformation data (16 data sets, 2 variables).
3. Data set 3: Clustering data set (14 data sets, 8 variables).
4. Data set 4: Reduction, transformation, and clustering data sets (14 data sets, 2 variables).

Results

In this section, the applications of SVR, ANN, and regression are first demonstrated. Then, comparisons are made to assess the efficacy of the employed methods.

SVR application

In support vector regression, the training and the validation data are specified first. The last three data sets (2004, 2008, and 2012) are specified as the validation data and the rest are designated for training purposes. To apply SVR, after dividing the data, parameters are specified. The radial of the acceptable cylinder for support vector regression is among these parameters and is denoted by ε. Unfortunately, a specific method for choosing the exact value of ε does not exist. Thus, ε, in this research, is specified by trial and error. Subsequently, the best interval for ε is specified as (0.01, 0.1). In addition, C, or the parameter for the loss function, is also specified through trial and error, and the optimal interval for C is specified as $(2^{-2}, 2^{4})$.

As mentioned earlier, Kernel functions are significantly influential functions in forecasting models. The radial basis Kernel function (RBF) is the utilized function in this research. RBF is the most common Kernel function that has been extremely beneficial in reducing error in models. Another advantage of RBF is that extensive parameter specification is un-obligatory for this function. The RBF only requires a single parameter to be specified. This parameter, Υ, is specified through trial and error in the range (2, 11). SVR is applied using the R software, based on which the following present the best models in the above-mentioned four different data sets:

I. Data set 1: the best model presents the parameters $C = 0.25$ and $\Upsilon = 2$, independent of ε.
II. Data set 2: the best model presents the parameters $C = 0.25$, $\Upsilon = 5$, and $\alpha = 0.07$.
III. Data set 3: the best model presents the parameters $C = 0.25$ and $\Upsilon = 2$.
IV. Data set 4: the best model presents the parameters $C = 0.25$, $\Upsilon = 5$, and $\alpha = 0.01$.

As the values in Table 5 demonstrate, the errors in the optimal models II and IV are less than the one in the model I. This clearly indicates how beneficial the preprocessing methods are in reducing the error of models. Moreover, the best model is identified as the model IV, which validates the benefit of using the clustering approach.

Table 5 SVR result for each data set

The optimal model of each data set	RMSE	MAPE	Improvement in RMSE (%)	Improvement in MAPE (%)
I	0.019546	3.544145	–	–
II	0.014997	2.43672	23.27	31.24
III	0.021612	4.432259	−10.56	−25.05
IV	0.010263	1.864497	47.49	47.39

ANN application

The artificial neural network is utilized and its results are further compared with the results obtained from SVR. To make an unbiased comparison, the data sets are the same as the previously four specified data sets. Furthermore, the last three sets are the designated as validation data, and the training data consist of the first 11 data sets from the clustered data.

To apply ANN, a multi-layer perceptron is chosen for the network as it has been highly successful in forecasting models. The architecture of the neural network involves the number of input and output neurons, the number of layers, the number of neurons in each layer, the connectivity of layers, and the transfer function in each layer. Furthermore, the number of input layers in each network is equal to the number of input variables, and the number of output layers is equal to the number of independent variables. There is no specific method to specify the number of neurons in hidden layers, and thus, trial and error are performed to specify this number. In general, the goal is to minimize the number of neurons within the hidden layers. This number is specified as 1 or 2 based on performed trials. Similar to the prior process, trial and error are utilized to specify the number of hidden layers. However, it should be considered that increasing the number of hidden layers will ultimately lead to an over-training situation. This, in turn, substantially increases the calculation time of the model. One or two hidden layers are utilized in the proposed model. Note that the number of estimated parameters must be less than the number of data sets.

Moreover, complete connectivity between the layers of the multi-layer perceptron network is considered, where each neuron in each layer is connected to all neurons in next layers. To reduce model complexity, a linear transfer function is specified in the output layer. The hyperbolic tangent and log-sigmoid function are utilized in other layers. Since the weight matrix is specified randomly at the beginning of the algorithm, the ANN procedure is applied more than once to obtain more accurate values for the weight matrix. Besides, a divergent network might be resulted due to inaccurate initial weights.

Table 6 Result of ANN for each data set

Model	RMSE	MAPE
I	0.015641	2.938875
II	0.030036	4.641433
III	0.01343	3.055326
IV	0.013959	2.586155

The architectures of the optimal ANN for the above-mentioned four data sets are as follows:

I. Data set 1: Number of hidden layers = 1, Number of neurons in each hidden layer = 1, Transfer function in each hidden layer is log-sigmoid.
II. Data set 2: Number of hidden layers = 1, Number of neurons in each hidden layer = 2, Transfer function in each hidden layer is log-sigmoid.
III. Data set 3: Number of hidden layers = 1, Number of neurons in each hidden layer = 1, Transfer function in each hidden layer is log-sigmoid.
IV. Data set 4: Number of hidden layers = 2, Number of neurons in each hidden layer = 1, Transfer function in each hidden layer is log-sigmoid.

Table 6 contains the performance measures of the above four ANNs. If *RMSE* is considered as the main performance measure, then models III and IV are identified as the best models. However, if *MAPE* is specified as the performance measure, then model IV is the best performing model. Ultimately, model IV is identified as the best ANN model.

Table 7 Result of linear regression for each data set

Model	RMSE	MAPE
I	0.129775	32.86371
II	0.084571	21.52493
III	0.141628	35.90823
IV	0.104456	26.22334

Table 8 Comparing the best model of each algorithm

Optimal model	RMSE	MAPE
SVR	0.010623	1.864497
ANN	0.013959	2.586155
Linear Regression	0.104456	26.22334

Linear regression

Linear regression is another utilized method that serves as a benchmark for the other algorithms. Table 7 demonstrates the results obtained using linear regression, based on which Model II is identified as the best performing model based on the calculated performance measures. Moreover, the values indicate that clustering is not beneficial in this case.

Comparison

To identify the final best forecasting model, the best performing model of each utilized method is selected and further compared based on the prior specified performance measures. Table 8 demonstrates the calculated *RMSE* and *MAPE* values associated with each method's best performing model.

The learning algorithms, SVR and ANN demonstrate lower values for *RMSE* and *MAPE* compared to Linear Regression. This indicates that the two learning algorithms outperform linear regression. In addition, the final best model is identified to be the SVR model, as its calculated *RMSE* and *MAPE* values are the lowest.

The SVR model is further applied to the data that have gone through preprocessing measures (clustering, data reduction, and transformation). The Kernel function of this model is RBF, and the parameter of this Kernel function is $\Upsilon = 5$. The other parameters are $C = 0.25$ and $\alpha = 0.1$. Table 9 demonstrates the predicted result of this approach:

The results in Table 9 indicate that the utilized SVR forecasting method is successful in forecasting the presidential election results in the last three elections. The number of necessary electoral votes to secure the presidency is set at 270. In both 2004 and 2012, where the incumbent party succeeds in the election, the predicted and the real electoral votes are higher than 270. However, in 2008 where the incumbent party is defeated, the predicted and the real electoral votes are significantly less than 270.

Table 9 Predicted and actual votes of the incumbent party

Election	Real electoral votes	Predicted electoral votes
2004	286	326.29
2008	173	173.54
2012	332	319.12

Conclusion and recommendation for future research

The objective of this research was to find an accurate forecasting model for the US presidential elections. Learning algorithms and data mining methods were utilized towards this objective. Moreover, independent variables such as GDP, unemployment rate, personal income, changes in the votes of the incumbent party in the last congress election, and the president's job approval were considered. The significance of each variable was determined by applying stepwise regression. Consequently, all variables except the president's job approval rate were omitted. The main theory that the presidential election is a referendum on the incumbent president's policies is proved to be true based on the findings. After the stepwise regression is performed, eight variables related to the president's job approval were considered to develop the forecasting model. By applying two preprocessing methods, data transformation and clustering, the data were prepared for the learning algorithms. Utilizing clustering and data transformation and reduction led to the accuracy of the model to improve by 50%. Furthermore, a comparison between the learning algorithms (SVR and ANN) and linear regression was carried out to identify the best model. The comparison demonstrated that the learning algorithms are by far better at reducing error compared to linear regression. Moreover, the SVR model was identified as the best performing forecasting model and proved successful in accurately forecasting the last three US presidential elections (2004, 2008, 2012).

In this paper, the variables were selected based on national statistics, but political and economic variables in each state are also significantly influential on people's decision in elections. Researchers could also take another approach and model the presidential elections in each state, and forecast based on the winner of each state. Besides, it is recommended to combine ANN and SVR with fuzzy systems, to improve forecasting accuracy. The major problem in applying SVR and ANN is a lack of a specific method to specify some parameters such as ε and Υ in SVR and the number of hidden layers in ANN. Researchers can further develop algorithms and heuristic methods that are capable of accurately specifying each method's necessary parameters.

References

Abramowitz AI (1988) An improved model for predicting presidential election outcomes. PS Polit Sci Polit 21:843–847

Abramowitz AI (2016) Will time for change mean time for Trump? PS Polit Sci Polit 49:659–660

Bashiri M, Farshbaf-Geranmayeh A, Mogouie H (2013) A neuro-data envelopment analysis approach for optimization of uncorrelated multiple response problems with smaller the better type controllable factors. J Indus Eng Int 9:30

Berg JE, Rietz TA (2014) Market design, manipulation, and accuracy in political prediction markets: lessons from the Iowa Electronic Markets. PS Polit Sci Polit 47:293–296

Charles CA, Reid GS (2016) Forecasting the 2016 general election in Jamaica. Commonw Comp Polit 54:449–477

Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20:273–297

De Neve J-E (2014) Ideological change and the economics of voting behavior in the US, 1920–2008. Elect Stud 34:27–38

Demuth H, Beale M (1998) Neural network toolbox: for use with MATLAB, 5th edn. The Math Works Inc, Natick

Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. In: Mozer MC, Jordan MI, Petsche T (eds) Advances in neural information processing systems 9. MIT Press, Cambridge, pp 155–161

Erikson RS, Wlezien C (2016) Forecasting the presidential vote with leading economic indicators and the polls. PS Polit Sci Polit 49:669–672

Fair RC (1978) The effect of economic events on votes for president. Rev Econ Stat 60:159–173

Fair R (2011) Predicting presidential elections and other things. Stanford University Press, Stanford

Ford R, Jennings W, Pickup M, Wlezien C (2016) From polls to votes to seats: forecasting the 2015 British general election. Elect Stud 41:244–249

Holbrook TM, DeSart JA (1999) Using state polls to forecast presidential election outcomes in the American states. Int J Forecast 15:137–142

Hosseini H, Luo D, Reynolds KJ (2006) The comparison of different feed forward neural network architectures for ECG signal diagnosis. Med Eng Phys 28:372–378

Karush W (1939) Minima of functions of several variables with inequalities as side constraints. Master's thesis, Dept. of Mathematics, Univ. of Chicago

Kuhn HW, Tucker AW (1951) Nonlinear programming. In: Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics, University of California Press, Berkeley, pp 481–492

Lewis-Beck MS (2005) Election forecasting: principles and practice. Br J Polit Int Relat 7:145–164

Lewis-Beck MS, Rice WT (1982) Presidential popularity and presidential vote. Public Opin Quart 46:534–537

Lewis-Beck MS, Rice TW (1992) Forecasting elections. CQ Press. Washington DC. http://works.bepress.com/tom_rice/4/. Accessed 15 Oct 2016

Lewis-Beck MS, Tien C (2014) Congressional election forecasting: structure-X models for 2014. PS Polit Sci Polit 47:782–785

Maleki MR, Amiri A, Mousavi SM (2015) Step change point estimation in the multivariate-attribute process variability using artificial neural networks and maximum likelihood estimation. J Indus Eng Int 11:505–515

Markopoulos AP, Georgiopoulos S, Manolakos DE (2016) On the use of back propagation and radial basis function neural networks in surface roughness prediction. J Indus Eng Int 12:389–400

Mattera D, Haykin S (1999) Support vector machines for dynamic reconstruction of a chaotic system. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in kernel methods—support vector learning. MIT Press, Cambridge, pp 211–242

Müller K-R, Smola AJ Rätsch G, Schölkopf B, Kohlmorgen J, Vapnik V (1997) Predicting time series with support vector machines. In: Gerstner W, Germond A, Hasler M, Nicoud JD (eds). Artificial Neural Networks—ICANN'97, Lecture Notes in Computer Science 1327, pp 999–1004

Munzert S (2017) Forecasting elections at the constituency level: a correction—combination procedure. Int J Forecast 33:467–481

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT press, Cambridge

Niaki STA, Hoseinzade S (2013) Forecasting S&P 500 index using artificial neural networks and design of experiments. J Indus Eng Int 9:1 **(9 pages)**

Rallings C, Thrasher M, Borisyuk G (2016) Forecasting the 2015 general election using aggregate local election data. Elect Stud 41:279–282

Saad EW, Prokhorov DV, Wunsch DC (1998) Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. IEEE Trans Neural Networks 9:1456–1470

Shokrollahpour E, Hosseinzadeh Lotfi F (2016) An integrated data envelopment analysis—artificial neural network approach for benchmarking of bank branches. J Indus Eng Int 12:137–143

Sigelman L (1979) Presidential popularity and presidential elections. Public Opin Quart 43:532–534

Stitson M, Gammerman A, Vapnik V, Vovk V, Watkins C, Weston J (1999) Support vector regression with ANOVA decomposition kernels. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in kernel methods—support vector learning. MIT Press, Cambridge, pp 285–292

Vapnik V (1982) Estimation of dependences based on empirical data. Springer, Berlin

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vapnik V, Chervonenkis A (1964) A note on one class of perceptrons. Automat Remote Control 25:821–837

Vapnik V, Chervonenkis A (1974) Theory of pattern recognition [in Russian]. Nauka, Moscow. (German Translation: Wapnik W. & Tscherwonenkis A., Theorie der Zeichenerkennung, Akademie-Verlag, Berlin, 1979)

Vapnik V, Lerner A (1963) Pattern recognition using generalized portrait method. Autom Remote Control 24:774–780

Wlezien C, Erikson RS (2004) The fundamentals, the polls, and the presidential vote. Polit Sci Polit 37:747–751

Yousefi M, Hooshyar D, Yousefi M, Khaksar W, Shahri KSM, Alnaimi FBI (2015) An artificial neural network hybrid with wavelet transform for short-term wind speed forecasting: a preliminary case study. 2015 International Conference on Science in Information Technology (ICSITech). IEEE. doi:10.1109/ICSITech.2015.7407784

Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14:35–62

Zhu X, Wang H, Xu L, Li L (2008) Predicting stock index increments by neural networks: the role of trading volume under different horizons. Expert Syst Appl 34:3043–3054