Forecasting US Presidential Election 2024 using multiple machine learning algorithms

Sinha, Pankaj and Kumar, Amit and Biswas, Sumana and Gupta, Chirag

Faculty of Management Studies, University of Delhi

20 October 2024

Forecasting US Presidential Election 2024 Using Multiple Machine Learning Algorithms

Pankaj Sinha, Amit Kumar, Chirag Gupta, Sumana Biswas

Faculty of Management Studies
University of Delhi

ABSTRACT

The outcome of the US presidential election is one of the most significant events that impacts trade, investment, and geopolitical policies on the global stage. It also sets the direction of the world economy and global politics for the next few years. Hence, it is of prime importance not just for the American population but also to shape the future well-being of the masses worldwide. Therefore, this study aims to forecast the popular vote share of the incumbent party candidate in the Presidential election of 2024. The study applies the regularization-based machine learning algorithm of Lasso to select the most important economic and non-economic indicators influencing the electorate. The variables identified by lasso were further used with lasso (regularization), random forest (bagging) and gradient boosting (boosting) techniques of machine learning to forecast the popular vote share of the incumbent party candidate in the 2024 US Presidential election. The findings suggest that June Gallup ratings, average Gallup ratings, scandal ratings, oil price indicator, unemployment indicator and crime rate impact the popular vote share of the incumbent party candidate. The prediction made by Lasso emerges as the most consistent estimate of the popular vote share forecast. The lasso-based prediction model forecasts that Kamala Harris, the Democratic Party candidate, will receive a popular vote share of 47.04% in the 2024 US Presidential Election.

1. INTRODUCTION

Elections worldwide are the true manifestation of people’s power of democracy in action. This process is crucial for the general population's participation in the shaping of governance processes and policies indirectly through the elected representatives and nation leaders. Amongst the most revered elected positions in the world, the post of United States (US) President is placed at the top rung of the political ladder. The person at the helm of affairs as the US president usurps a substantial amount of power not just on the national but on the world stage too. The US president has a large say in determining global policies and thus steers the direction of economic and geopolitical developments at the global level. The US, being the largest economy, confers a significant amount of influence on global trade, investment, exchange rates, inflation, growth rates, etc. The results of the US Presidential elections thus assume significance, especially in the current scenario where the world is grappling with the severe impact of the COVID-19 pandemic. All the more, the result of the upcoming US Presidential elections holds a lot more significance as the geo-political tensions have heightened in different parts of the Asian continent in the form of the long-stretched Russia-Ukraine war, the sudden increase in the level of direct conflict in the middle-east and the rise of Chinese territorial claims in the South China Sea region along with its stance on the sovereignty of Taiwan. The prediction of the upcoming 2024 US Presidential election results has become a key issue as the two candidates contesting the election have very different views on national as well as global issues. The prediction of the 2024 US Presidential elections is based on the key factors influencing the voting behaviour of the US population. The voting pattern is expected to be influenced by numerous economic and non-economic issues at the center stage of this and previous elections. Voters face various macro-level and micro-level factors that affect the difficulty in choosing candidates to contest in the elections. The complexity of the political information environment often further complicates the decision-making by voters and, thus, the prediction of election results (Lau et al., 2008). Therefore, political heuristics should be applied to simplify the decision regarding the contesting candidates. People use certain broad and inclusive factors as shortcuts to form an opinion regarding the Presidential candidates rather than processing a substantial amount of information on multiple issues. One of the crucial political heuristics is the way the incumbent party has managed the economy, which is reflected in the GDP growth rate, level of inflation, strength of the currency in the international markets, etc., during the last regime in comparison to the time when the opposition was in power (Owens, 1984). The perceived performance of the incumbent party takes center stage in the decision-making (Shaffer & Chressanthis, 1991). The crime rate, unemployment rate, burden of taxes on the highly influential corporate sector, scandalous or non-scandalous nature of the incumbent President, and campaign spending also play a crucial role in forming various perceptions in the voters' minds that eventually determine the election results. The perceived performance scores of the incumbent president reflected in the ratings published by Gallup and through the mid-term performance do have a substantial say in determining the voting pattern in the elections. The impact of various economic and non-economic fundamentals on US elections pushes a large population to vote beyond the party lines or affiliations (Erikson & Wlezein, 2014). It was also seen that the issues and leadership model that relies on how the people perceive the candidates' ability to manage important issues and how well they exhibit leadership traits perform well in predicting the votes to be received by the Presidential candidates in the US elections (Graefe, 2013). It was also observed that party identification takes a backstage in the choice of candidate as the campaign progresses, with key economic and non-economic issues coming to the forefront in determining voting behaviour (Graefe, 2013). A comprehensive analysis of how economic conditions under the incumbent president shape the voting behaviour in the US revealed that the economic performance of the incumbent president has formed the very base of voting decisions in the US since 1790 (Guntermann, 2021). It also plays a key role in the incumbent's decision on whether to contest again (Guntermann, 2021). A broader variable of war was included amongst the other macro indicators to develop a more generalised model for election results forecasting (Walker, 2006). It has also been observed that economic performance manifested in GDP growth, unemployment and exchange rate numbers do impact the US Presidential election results and are key factors in forecasting the same (Sinha et al., 2020; Sinha et al., 2024). Another study also found that voters objectively assess the economic performance of the incumbent president, going beyond party identification, which acts as a key factor in their voting decision (Cambell, 2005; Holbrook, 2009; Lewis-Beck & Martini, 2020). Apart from the incumbent president's performance on the key economic fundamentals, the knowledge about the views and standing of the contesting candidates on the key issues in each election significantly influences the outcome of the elections (Gelman & King, 1993). Some key issues in the upcoming election are illegal immigration, corporate tax cuts, etc., which effectively reach the voters through media channels and form the foundation of the enlightenment preference hypothesis in the US elections (Gelman & King, 1993). It was suggested that social issues had increasingly mattered to the voters in the US Presidential elections between 1992-2012, and a liberal outlook of the Democratic party gives them an edge on that front over the central idea of a higher defence budget and government provisioning highlighted by the Republican Party (Wurgler & Brooks, 2014). It was further observed that non-economic factors played a bigger role in influencing the outcome of US Presidential elections (Sinha et al., 2016). Another interesting finding revealed from 1960 to 2000 was that a higher voter turnout was beneficial for the Democratic Party in the US elections (Martinez & Gill, 2005). However, the bread and peace model emphasised that the count of US soldiers who lost their lives to wars and the extent of change in the per capita real disposable income that determines the consumption capability of an individual strongly determines the chances of winning the election by the incumbent party (Hibbs, 2000). The economic and non-economic indicators' influence on the election results are further moderated by the cost of being in power for a longer duration as it leads to a drift in the policy stance of the incumbent president over a longer duration, which does not sink in well with the voters' expectations (Cambell, 2005; Wlezein, 2017). Another significant factor in recent times that impacts election outcomes is campaign spending, which was found to significantly influence the voters' decisions, especially less-informed voters and voters who are highly critical of the economic performance of the incumbent president (Erikson & Palfrey, 1998; Jacobson, 2006; Schuster, 2020). The aspect of campaign spending becomes highly crucial in closely fought contests as it plays a significant role in increasing vote share in the elections in the US (Nagler & Leighley, 1992). It is also important to note that incumbents embroiled in scandals experience a thinning of their vote share as scandals create a negative image of the incumbent and thus become an important decisive factor in the US Presidential elections (Praino et al., 2013; Sinha et al., 2020; Sinha et al., 2024). Incumbent candidates or parties become more vulnerable to the negative effect of scandals in the presence of lower approval by voters identified with the opposition party and lower traffic in the news items, which enhances the intensity of coverage by media channels (Nyhan, 2015; Von Sikorski, 2018; Von Sikorski et al., 2020). However, it was further suggested that scandal-plagued candidates may be able to weather off the ill effects by sharing a positive stance on other important issues, the presence of strong partisan loyalties among voters and other positive news about the characteristics of the candidate as voters do not completely rely on the moral lens while judging a candidate’s abilities to hold the position of the US president (Funck & McCabe, 2022). A comprehensive analysis of scandals over time revealed that the negative effects varied across time and contexts, raising a pertinent question of the efficacy of scandals to impact election results (Rottinghaus, 2023).

The above discussion elaborates on the impact of various economic and non-economic indicators on voting behaviour in the US Presidential elections. It is observed that both economic and non-economic factors are at play and should be considered while modelling the forecast of the upcoming 2024 US Presidential elections. This study extends the work of Sinha et al. (2024) by employing machine learning algorithms. The purpose of employing machine learning techniques is to develop a more generalised prediction model by capitalising on the ability of machine learning algorithms to attain lower bias and variance simultaneously. This helps in arriving at a more accurate forecast of the popular vote share of the incumbent party in the forthcoming 2024 US Presidential elections. This study includes all the economic and non-economic factors used in Sinha et al. (2024) in the initial analysis to select the most parsimonious set of variables impacting Presidential elections. Later, it uses these variables with different techniques to get the popular vote share forecast. Refer to Sinha et al. (2024) for variable names and explanations.

2. METHODOLOGY

The forecast of the popular vote share for the incumbent party is generated by applying machine learning techniques following a two-step process. Firstly, this study selects the sparse set of features or variables from the group of economic and non-economic indicators used by Sinha et al. (2024) that predicted the popular vote share using a linear regression method. The feature or variable selection is done using a regularization-based machine learning algorithm. Secondly, the variables selected in the first step are used as inputs in three different types of machine learning techniques to forecast the vote share. Popular vote share is the dependent variable. The explanatory variables consist of both economic and non-economic factors used in Sinha et al. (2024). The dataset for the study covers the period from 1952-2020. The data set's time interval is four years, corresponding to the gap between two election cycles. The values of each variable used in the study are shown in the Sinha et al (2024) appendix. The list of variables, along with their identification codes, is presented in Table 1.

Table 1- Identification codes for variables

Sl. No.	Variables	Identification code
A	**Dependent Variable**	
	Popular vote share	PVOTE
B	**Independent variable**	
	Economic variables	
1	Unemployment rate	Unemploy1
2	Inflation	Inflation1

Sl. No.	Variables	Identification code
3	Economic Growth Rate	GDPR
4	Gold Prices	Gold_Return
5	Gold Price Index	Gold_index
6	Exchange Rate	Exchange1
7	Oil Prices	Oil1
8	Corporate Tax Rate	Tax_R1
9	Corporate Tax Impact	Tax
	Non-Economic Variables	
1	Gallup Job Approval Rating	Jun_Gal
2	Average Gallup Rating	Avg_Gal
3	Crime Rate	Crime_R
4	Mid-Term Performance	Mid_Term
5	Period of Power	Power
6	Campaign Spending Index	Spending
7	Incumbent President Running	Incum_PRun
8	Scandals	Scandals
9	Illegal immigration	Illegal_Immig
10	Illegal Aliens	Illegal_Aliens

Note: Refer to Sinha et al. (2024) for an explanation of the variables in Table 1.

The empirical model for forecasting the popular vote share is represented by Equation 1.

$$PVOTE_t = \beta_0 + \sum_{i}^{t} c_i X_i + \varepsilon_t \qquad (1)$$

$PVOTE_t$ in equation 1 represents the popular vote share percentage of the incumbent party. X_i denotes all the independent variables listed in Table 1. ε_t is the error term in year t.

Traditional statistical methods, such as linear regression (OLS), are unable to capture the proper relationship among underlying variables, and this is called bias. The best-fit line is a straight line; in machine learning, the ideal algorithm has low bias and can accurately model or capture the true relationship, and it also has low variability as it can give consistent prediction results across different datasets or on unseen data. Thus, machine learning algorithms have at times proved to be more powerful than linear regression analysis done traditionally for forecasting models, depending on the nature of the data and forecast. The non-linear relationship in data is especially better captured through such algorithms. It is also seen that machine learning models better deal with data involving many independent variables and their interactions in the model as they detect the relationship and interaction between features more robustly. The generalised model is obtained by finding the optimal spot between the simple and complex models. The three standard methods to find the optimal model are regularisation, bagging and boosting.

The regularization technique is used to prevent overfitting and underfitting by adding a penalty to the loss function in statistical modelling and machine learning. Overfitting occurs when a model learns the underlying patterns in the training data, including noise, to the extent that it does not perform well on new data. Regularisation helps to improve model generalisation by constraining the complexity of the model.

Bagging, or Bootstrap Aggregating, is a technique used in Random Forest to reduce variance and improve the model's accuracy. In bagging, several decision trees are trained independently on different bootstrap samples (random subsets) of the training data. Each tree generates a prediction, which is averaged (in regression) or voted upon (in classification) to produce the final output. Each tree is trained on a slightly different data set, allowing the Random Forest to generalise better to unseen or new data. This ensemble method improves model robustness and accuracy compared to using a single decision tree.

Boosting, on the other hand, is a sequential ensemble technique employed by Gradient Boosting to improve model accuracy by focusing on the mistakes of the previous models. Each weak learner (usually a shallow decision tree) is trained to correct the residual errors of the prior model, iteratively improving the overall prediction. Boosting minimises the loss function, such as the Mean Squared Error (MSE) for regression, by adding models with higher complexity in a step-by-step manner.

This study uses all three processes of regularisation, bagging and boosting to arrive at the popular vote share forecast. Three machine learning models incorporating the earlier mentioned methods are used in the present study to predict the popular vote share of current party candidates in the upcoming 2024 election. These techniques are LASSO (Least Absolute Shrinkage and Selection Operator), Random Forest (RF) and Gradient Boosting (GB). Lasso was first used to identify the most significant predictor variables from the set of both economic and non-economic factors used in Sinha et al. (2024). It is also later used to forecast the popular vote share of the incumbent party candidate for the 2024 US Presidential election. RF and GB also use the lasso-identified variables to predict the popular vote share forecast of the incumbent party candidate for the 2024 US Presidential election.

LASSO (LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR)

Lasso performs multiple iterations of regression with different combinations of variables to determine the most sparse set of features necessary for prediction. It employs regularisation to estimate a generalised model with low bias and low variance, thereby improving the model's interpretability and forecast accuracy. The Lasso technique uses a penalty to shrink some of the regression coefficients to zero. Doing this retains the significant features in the model, and other non-significant ones are rejected. This results in a more illustrative model. The objective function representing the Lasso model is given in Equation 2.

$$\min_{\beta}\left(\sum_{i=1}^{n}(y_j - X_i\beta)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|\right) \qquad (2)$$

Here Yi is the actual observed value for the i-th sample, $X_i\beta$is the predicted value from the linear model, β_j are the model coefficients. λ is the regularization parameter and it controls the amount of shrinkage which is applied to the coefficients. The penalty term $\lambda\sum_{j=1}^{p}|\beta_j|$ is the sum of the absolute values of the coefficients. As λ increases, the penalty increases, causing some coefficients to shrink to zero.

Lasso starts with a basic regression model. Let's consider we want to predict a target y_i for each observation i, based on a set of features X. The initial prediction is based on a linear combination of the features shown in equation 3.

$$\hat{y} = \beta_0 + \sum_{j=1}^{p}\beta_j x_j \qquad (3)$$

where β_0 is the intercept and β_j are the coefficients corresponding to the features x_j.

Lasso minimizes the combination of the residual sum of squares and the L_1 penalty term. As the value of λ grows, the model eliminates non-significant features by setting their coefficients to zero, achieving both feature selection and regularization as shown in equation 4.

$$L(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \qquad (4)$$

When λ=0, the Lasso model becomes similar to ordinary least squares (OLS) regression without applying regularization. The value of λ controls the trade-off between fitting the model to the training data and enforcing sparsity by shrinking the coefficients.

RANDOM FOREST (RF)

Regression by Random Forest is a useful machine learning technique which builds and combines multiple decision trees during classification and regression to reach to a final model prediction. It reduces data overfitting and thus improves accuracy. The model combines high variance decision trees and reach to a resultant low variance output. For the classification purpose the model uses the majority voting classifier for the final output. The key idea is to aggregate the predictions of multiple trees to form a robust model. While in a regression problem, the mean of all the predicted values given by the all the trees is considered to be the final output.

The prediction for regression using this model is represented in Equation 5.

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t \qquad (5)$$

Where, T denotes the number of decision trees and $\hat{y}_t$ is the estimate of the t-th tree.

Random Forest randomly creates bootstrap samples for each decision tree which involves repeated random sampling by selecting some data point multiple times, while not selecting others at all. For example, for a data set with n observations shown in equation 6

$$\{x_1 , x_2 \; x_3 \; x_4 , \dots\dots\dots , x_n \;\} \qquad (6)$$

Bootstrap sample example is represented in equation 7.

$$\{x_4 , x_8 \; x_3 \; x_4 , x_7 , \dots\dots , x_n , x_{n-5} , x_2 , \; x_8 \dots\dots\} \qquad (7)$$

Some observations in equation 7 appear more than once (e.g., x_4 , x_8) while others do not appear (e.g., x_1 , x_5). At each node of a decision tree, to determine the best split, a random subset of features is selected. This introduces further randomness and helps to de-correlate the trees, reducing overfitting. The number of features considered for each split is a hyperparameter.

For classification purpose, majority voting is used by Random Forest, where each tree "votes" for the forecasted class, and the class having the most number of votes is selected. Random Forest uses bagging (Bootstrap Aggregating) to combine the predictions of multiple trees, reducing variance and preventing overfitting shown in equation 8.

$$\hat{y} = mode\;(y_1, y_2 \dots . \, y_T) \qquad (8)$$

At each split, only a subset of features is considered, introducing diversity among the trees and making the forest more robust to noisy data.

GRADIENT BOOSTING (GB)

Gradient Boosting is a machine learning algorithm primarily used for regression and classification tasks. In this technique, models are built sequentially, and each new model aims to correct the errors made by the previous models. The method uses decision trees as weak learners and combines them to produce a strong predictive model. The core concept behind Gradient Boosting is to minimize the loss function by iteratively improving the model, making it suitable for capturing complex patterns in data. Unlike Random Forest, which builds trees independently, Gradient Boosting builds trees that learn from the mistakes of the previous trees by leveraging gradient descent optimization.

The process begins by fitting a simple model (often a constant value like the mean) to the data. To predict a target y_i for each observation i, we start with an initial prediction $F_0(x)$, which can be the mean of the target values in case of regression shown in equation 9.

$$F_0(x) = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad (9)$$

At each iteration m, a new weak learner (often a decision tree) is fitted to the residuals or pseudo-residuals, which are the differences between the actual values y_i and the current prediction $F_m(x_i)$. The new tree $h_m(x)$ is trained to predict these residuals, and the current model is updated as shown in equation 10.

$$F_{m+1}(x_i) = F_m(x_i) + \alpha \cdot h_m(x_i) \qquad (10)$$

Here, α is the learning rate, which controls the contribution of each tree, and $h_m(x_i)$ represents the new weak learner at iteration m.

To minimize the loss function, Gradient Boosting uses gradient descent technique. The residuals can be thought of as the negative gradient of the loss function with respect to the predictions. For a regression task with the mean squared error (MSE) as the loss function, the gradient at iteration m for each data point i is given by equation 11.

$$g_{\mathrm{m},i} = -\frac{\delta L\big(y_i, F_m(x_i)\big)}{\delta F_m(x_i)} = -(y_i - \ F_m(x_i)) \qquad (11)$$

After a certain number of iterations M, the final model combines all the weak learners. The final prediction for each observation is given by equation 12.

$$\hat{y} = F_M(x) = F_0(x) + \sum_{m=1}^{M} \alpha \cdot h_m(x) \qquad (12)$$

The contribution of each tree is scaled by the learning rate α, which ensures that the model does not overfit the data by taking large steps at each iteration. The learning rate controls the amount of contribution by each weak learner to the overall model. A smaller learning rate requires more iterations (trees), but typically leads to better generalization. The number of boosting rounds or iterations M controls how many weak learners are added. A higher number of trees generally improves the model but can lead to overfitting if not properly controlled by other regularization techniques.

Thus, this study applies regularization to select the predictor variables from the given set and further applies regularization, bagging and boosting to forecast the popular vote share of the incumbent party. Grid search is applied to find out the optimal value of hyperparameters for all the three machine learning techniques.

3. EMPIRICAL RESULTS

This study modifies the feature selection technique in comparison to the method used by Sinha et al (2024). The regularization process is applied by using the lasso technique to select the most important variables impacting the voting pattern in comparison to the stepwise backward elimination and intuitive, iterative process used by Sinha et al. (2024). Lasso selects June Gallup approval ratings, Average Gallup ratings, scandal ratings, level of unemployment, oil prices and crime rate as the most important variables impacting the popular vote share of the incumbent party candidate. These six variables, which consist of factors from both economic and non-economic categories, are further incorporated with lasso, RF and GB machine learning algorithms to forecast the popular vote share of the incumbent party in the forthcoming 2024 US Presidential elections. The final proposed model for the popular vote share forecast is given in Equation 13.

$$\mathbf{PVOTE = C + C1Jun\_Gal + C2Unemploy1 + C3Oil1 + C4Crime\_R + C5Scandals + C6Avg\_Gal +} \boldsymbol{\varepsilon} \qquad \mathbf{(13)}$$

Table 2 presents the results obtained for the popular vote share percentage prediction obtained from the three machine learning techniques of lasso, random forest and gradient boosting. Table 2 provides the coefficient values for all the variables for lasso and the feature importance values given by random forest and gradient boosting. The r squared value, best hyperparameter value obtained from grid search along with the final predicted popular vote share of the incumbent party in 2024 elections is also shown in table 2. The proposed model with six variables shown in equation 13 is used with lasso, RF and GB to back test the efficiency of all the three techniques.

Table 2: Empirical results of popular vote share forecast of incumbent party for 2024 Presidential election

Prediction Results using Variables selected by Regularization			
Variables	**LASSO Model**	**Random Forest Regressor**	**Gradient Boosting Regressor**
Jun_Gal	0.006603	0.5076	0.6187
Unemploy1	0.004345	0.1162	0.1396
Oil1	0.000468	0.0456	0.009
Crime_R	0.000004	0.0715	0.1141
Scandals	-0.001996	0.1754	0.1104
Avg_Gal	-0.003274	0.0836	0.0083
Intercept	0.331820	-	-
Hyperparameter	Alpha = 0.01	Best N_Estimators: 32	Best N Estimators: 10, Best Learning Rate: 0.9899, Best Max Depth: 6
R^2	0.7697	0.9160	1
Forecasted Popular Vote Share Percentage 2024	47.04%	46.62%	46.79%

Table 3: Backtesting results for 2012, 2016 and 2020 US Presidential elections

Backtesting for US Presidential Election						
(Variables selected by Regularization)						
Year	**LASSO Model**		**Random Forest**		**Gradient Boosting**	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
2012	51.01%	49.89%	51.01%	54.23%	51.01%	50.83%
2016	48.02%	47.71%	48.02%	56.34%	48.02%	58.77%
2020	46.80%	42.64%	46.80%	44.17%	46.80%	42.72%

Backtesting performance for all three previous elections shows that Lasso exhibits better consistency and prediction accuracy for two out of three years, whereas RF and GB performed well in only one out of three previous elections. Lasso gives an error of 3.22% (2012), 8.32% (2016) and 2.63% (2020), with an average error of 1.86% in forecasting the popular vote share of the party. RF gives an error of 3.22% (2012), 8.32% (2016) and 2.63% (2020) with an average error of 4.72%, whereas GB gives an error of 0.18% (2012), 10.75% (2016) and 4.08% (2020) with an average error of 5%. Therefore, lasso provides more consistent and accurate estimates of the popular vote share for the previous elections. Backtesting or predicting the popular vote share for the previous three elections involves the processing of data for all the years from the beginning till the year (election year) before the year for which the estimation is done, that is, while estimating the popular vote share for 2012, the data till 2008 is used and similarly for 2016 and 2020 estimations data till 2012 and 2016 is used respectively.

Empirical results in Table 2 show that the R-squared value is 76.97%, 91.6% and 100% for lasso, RF and GB. The very high r squared value given by RF and a perfect 100% r square given by GB give the impression that these two techniques are more accurate than lasso. However, it is not reflected so in the backtesting results. This anomaly is because RF and GB do not perform well with a small dataset with fewer features. The dataset used in for prediction has only 17 rows and 6 variables or features. Both RF and GB are highly susceptible to overfitting with small datasets as these frameworks are built to learn complex patterns exhibited in high dimensional large datasets. They tend to overfit the small datasets as they may completely memorize the training data and thus fail to build generalized models for handling unseen data. This results in high variance as they perform extremely well on training data but end up with inaccurate estimates for unseen data. High variance means that models change drastically with a very small change in the data, thereby making unreliable models inappropriate for generalisation. Limited data reduces the efficiency of ensembling techniques such as RF and GB. The small nature of our data also creates a problem of limited diversity. Thus, RF and GB cannot explore robust patterns, which also creates a problem in tuning the hyperparameters in these techniques. This is visible in the high R squared value of RF and GB. However, low accuracy or inconsistent estimates in backtesting are visible in high average prediction errors for the previous three US presidential elections.

Thus, lasso emerges as the most appropriate technique given the peculiar characteristics of the data used in the study. The popular vote share of the incumbent party candidate in the 2024 US Presidential Elections is forecasted using the proposed model in Equation 13. The actual values of the six predictor variables identified by lasso for 2024, asgiven in Table 14, have been used with the coefficients obtained by lasso.

Table 4- Value of independent variables from the year 2024

Independent variables	2024_Values
Jun_Gal	38
Unemploy1	0.37
Oil1	40.63
Crime_R	1960.95
Scandals	0
Avg_Gal	43

The forecasted vote share of the Incumbent Democratic Party presidential candidate Kamala Harris is 47.04%.

The study applied regularisation to select the predictor variables to be used with regularisation (lasso technique), bagging or bootstrap aggregation (random forest), and boosting (gradient boosting) to forecast the popular vote share percentage. The final proposed model comprised of both economic and non-economic factors. Lasso was selected as the best model, which forecasted the popular vote share percentage of Kamala Harris (incumbent party presidential candidate) to be 47.04% in the upcoming 2024 US Presidential elections.

4. CONCLUSION

This study is an extension of Sinha et al. (2024) as it starts with the same group of economic indicators and non-economic indicators to build a prediction model for forecasting the popular vote share percentage expected to be receivedby the Democratic (incumbent) party presidential candidate, Kamala Harris, in the forthcoming US Presidential election. This study changes the methodology by adopting the three machine learning techniques: regularisation, bagging and boosting. These processes are particular frameworks built to find the optimal model for achieving greater generalisation in prediction accuracy through the bias-variance trade-off. The study employs three machine learning algorithms to arrive at the popular vote share forecast: lasso, random forest, and gradient boosting. Among the independent variables, June Gallup ratings, average Gallup ratings, level of unemployment, oil price indicator, crime rate indicator and scandal rating are selected by Lasso as the most important forces influencing the electorate's voting behaviour. The popular vote share forecast of Kamala Harris is 47.04%, given by Lasso, which emerged as the most consistent technique with low bias and low variance out of the three machine learning methods used for forecasting.

REFERENCES

- Campbell, J. E. (2005). The fundamentals in US presidential elections: Public opinion, the economy and incumbency in the 2004 presidential election. *Journal of Elections, Public Opinion & Parties*, *15*(1), 73-83.
- Erikson, R. S., & Palfrey, T. R. (1998). Campaign spending and incumbency: An alternative simultaneous equations approach. *The Journal of Politics*, *60*(2), 355-373.
- Erikson, R. S., & Wlezien, C. (2014). Forecasting US presidential elections using economic and noneconomic fundamentals. *PS: Political Science & Politics*, *47*(2), 313-316.
- Funck, A. S., & McCabe, K. T. (2022). Partisanship, information, and the conditional effects of scandal on voting decisions. *Political Behavior*, *44*(3), 1389-1409.
- Gelman, A., & King, G. (1993). Why are American presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, *23*(4), 409-451.
- Graefe, A. (2013). Issue and leader voting in US presidential elections. *Electoral Studies*, *32*(4), 644-657.
- Guntermann, E., Lenz, G. S., & Myers, J. R. (2021). The impact of the economy on presidential elections throughout US history. *Political Behavior*, *43*, 837-857.
- Hibbs Jr, D. A. (2000). Bread and peace voting in US presidential elections. *Public Choice*, *104*(1), 149-180.
- Holbrook, T. M. (2009). Economic considerations and the 2008 presidential election. *PS: Political Science & Politics*, *42*(3), 473-478.
- Jacobson, G. C. (2006). Campaign spending effects in US Senate elections: evidence from the National Annenberg Election Survey. *Electoral Studies*, *25*(2), 195-226.
- Lau, R. R., Andersen, D. J., & Redlawsk, D. P. (2008). An exploration of correct voting in recent US presidential elections. *American Journal of Political Science*, *52*(2), 395-411.
- Lewis-Beck, C., & Martini, N. F. (2020). Economic perceptions and voting behaviour in US presidential elections. *Research & Politics*, *7*(4), 2053168020972811.
- Martinez, M. D., & Gill, J. (2005). The effects of turnout on partisan outcomes in US presidential elections 1960–2000. *The Journal of Politics*, *67*(4), 1248-1274.
- Nagler, J., & Leighley, J. (1992). Presidential campaign expenditures: Evidence on allocations and effects. *Public Choice*, *73*(3), 319-333.
- Nyhan, B. (2015). Scandal potential: How political context and news congestion affect the president's vulnerability to media scandal. *British Journal of Political Science*, *45*(2), 435-466.

- Owens, J. R. (1984). Economic influences on elections to the US Congress. *Legislative Studies Quarterly*, 123-150.
- Praino, R., Stockemer, D., & Moscardelli, V. G. (2013). The lingering effect of scandals in congressional elections: Incumbents, challengers, and voters. *Social Science Quarterly*, *94*(4), 1045-1061.
- Rottinghaus, B. (2023). Do Scandals Matter? *Political Research Quarterly*, *76*(4), 1932-1943.
- Schuster, S. S. (2020). Does campaign spending affect election outcomes? New evidence from transaction-level disbursement data. *The Journal of Politics*, *82*(4), 1502-1515.
- Shaffer, S. D., & Chressanthis, G. A. (1991). Accountability and US Senate elections: A multivariate analysis. *Western Political Quarterly*, *44*(3), 625-639.

- Sinha, Pankaj & Verma, Kaushal & Biswas, Sumana & Tyagi, Shashank & Gogia, Shaily & Singh, Aakhyat & Kumar, Amit, 2024. "**Modeling and forecasting US presidential election 2024**," MPRA Paper 122319, University Library of Munich, Germany, revised 08 Oct 2024.

- Sinha, Pankaj & Srinivas, Sandeep & Paul, Anik & Chaudhari, Gunjan, 2016. "**Forecasting 2016 US Presidential Elections Using Factor Analysis and Regression Model**," MPRA Paper 74618, University Library of Munich, Germany, revised 17 Oct 2016.

- Sinha, P., Verma, A., Shah, P., Singh, J., & Panwar, U. (2022). Prediction for the 2020 United States presidential election using machine learning algorithm: Lasso regression. *Journal of Prediction Markets*, *16*(1).
- Von Sikorski, C. (2018). Political scandals as a democratic challenge| The aftermath of political scandals: A meta-analysis. *International journal of communication*, *12*, 25.
- Von Sikorski, C., Heiss, R., & Matthes, J. (2020). How political scandals affect the electorate. Tracing the eroding and spillover effects of scandals with a panel study. *Political Psychology*, *41*(3), 549-568.
- Walker, D. A. (2006). Predicting presidential election results. *Applied Economics*, *38*(5), 483-490.
- Wlezien, C. (2017). Policy (mis) representation and the cost of ruling: US presidential elections in comparative perspective. *Comparative Political Studies*, *50*(6), 711-738.
- Wurgler, E., & Brooks, C. (2014). Out of step? Voters and social issues in US presidential elections. *The Sociological Quarterly*, *55*(4), 683-704.