Assessing the Impacting Factors in Prediction of Parliamentary Elections Turnout Using Heuristics and Devising MIP Algorithmic Model

Lavdim Beqiri
Computer Sciences
University Goce Delcev, Shtip
Shtip, North Macedonia
lavdim.beqiri@gmail.com

Majlinda Fetaji
Computer Sciences
South East European University
Tetovo, North Macedonia
m.fetaji@seeu.edu.mk

Bekim Fetaji
Informatics, Computer Sciences
Mother Teresa University
Skopje, North Macedonia
Bekim.Fetaji@unt.edu.mk

Zoran Zdravev
Computer Sciences
University Goce Delcev, Shtip
Shtip, North Macedonia
Zoran.zdravev@ugd.edu.mk

***Abstract*— The research focus is on identifying and assessing the impacting factors to be measured in order to realize prediction of parliamentary elections outcome using the (MIP) algorithmic model. We have developed a novel method for recognizing the main impacting factors in elections using the (MIP) algorithmic model. We have firstly used adaptive heuristics. In order to devise and asses the impacting factors we have devised most-important-problem (MIP) algorithmic model to predict the outcome of Kosovo parliamentary elections and grounded it on the TTB (take-the-best) strategy. An analysis of forecasting approach to elections and the performance metrics (variance) using the (MIP) algorithmic model has been used. provided are all the main variables we have measured. We have provided posterior binomial proportion. This method is very popular when modelling geopolitical situations with complex dynamics in the system. The data has derived from an originally collected survey dataset that contains the impacting factors previously identified and assessed regarding the parliamentary elections in Kosovo has been realized.**

Keywords*— *parliamentary election turnout prediction; heuristic; machine learning; (MIP) algorithmic model

I. Introduction

There are usually a number of information sources we can turn to when we need to make a decision in the midst of ambiguity. One such typical situation is prediction of elections then there are uusually a variety of information sources to turn to. The main problems we face today are unpredictable elections, the result of which is detrimental to our democracy. What variables as impacting factors to be measured should be investigated for election prediction? Because of the lack of analytical tools to simulate voting trends, we have developed a novel method for recognizing the main impacting factors in parliamentary elections. We have firstly used heuristics. Two of the most commonly used variables are income and education. Other factors include party affiliation, age, and gender. All these things have a tendency to predict how someone will vote. We also introduced another variable probable-winner-is, so the voters can do their prediction and include that variable also in our model. So, our major contribution is the creation of an innovative algorithm extending the processing of information produced by the learning algorithms, to improve mining and prediction of elections..

II. Purpose and Aim of Study

The study is aiming to summarize the knowledge about prediction of elections and characteristics in terms of the strength (algorithm strength) with respect to voter characteristics (political party affiliation and their regional characteristics) which are collected from their data collected during the 2021 general elections in Kosovo. Specifically, we are interested in studying the prediction of elections with our new, innovative algorithmic model and later using deep learning machine learning algorithm. We begin our analysis by considering variables that tend to predict strength with respect to political party affiliation. We summarize. our findings in a table entitled, “Election Prediction”. We start by examining the reasons for election failure. This is not just the absence of a voter, but the absence of a desire to vote. We analysed data sets for voting patterns run on Google Trends for 2008 to 2021.We added information regarding the political ideology of voters, and we determined the 48 prerequisites for successful elections: if there are more committed voters than disinterested voters, the politician that is supported by the majority of the voters will emerge victorious. However, recent elections and political history are replete with controversial results: the date May 11, 2008 year, marked the transition from one-party state to a multi-party one, and again from one political actor winning to the opposition party emerging victorious.

III. Research Methods

We have used the Adaptive Heuristic Toolbox and devised the most-important-problem (MIP) algorithmic model to predict the outcome of Kosovo parliamentary elections. The adaptive heuristic toolbox is the collection of guidelines or heuristics it has access to at a certain stage of its evolution as discussed by [1]. The TTB (take-the-best) approach is one of the most fundamental algorithms in the adaptive heuristic toolbox and the subject of this article. Following guidelines from [1] asserted that we choose and use a range of quick and economical heuristics from our cognitive "adaptive toolbox" when making judgments. Take-the-best (TTB), developed by [2], is a heuristic for selecting between options using just one piece of information. That is, a decision-maker examines each quality, starting with the most crucial one, to see if it distinguishes between the options. If the answer is affirmative, he chooses the course of action that the attribute favors. If not, he moves on to the subsequent most crucial quality. TTB was very predictive, according to [4]. For 20 prediction problems, the number of variables ranged from 3 to 19, and the authors compared the heuristic to multiple regression and unit-weighting. These problems included forecasting high school dropout rates, male and female attractiveness, homelessness and mortality in American cities, college professor salaries, childhood obesity, and fish fertility. The majority of these instances have a goal to illustrate the use of multiple regression analysis in statistics textbooks. Unsurprisingly, when forecasts were calculated using samples, multiple regression fared best. TTB, however, was the most precise, followed by unit-weighting, when using cross-validation to forecast data that had not been used to train the model. When there were fewer observations for each predictor variable, TTB had a greater advantage. Multiple regression rarely outperforms TTB, even when there are more than 10 data for each variable. [4] found additional requirements for the TTB's applicability. They demonstrated analytically that if the implied importance weight of a variable exceeds the total weights of all less important variables, a linear model cannot outperform TTB. The findings showed that TTB use was more common in predictable environments, when information costs were high, and when the validity of the cues was recognized.

We developed the most-important-problem (MIP) voting model to predict the outcome of upcoming Kosovo parliamentary elections. The model is based on data on voters' expectations for how candidates would approach the problem they believe to be most significant. The winner of the popular vote is determined by the model using only polling data and a heuristic like TTB. For the year leading up to election year, we gathered polling data on the issue respondents believe is the most crucial one facing the nation (e.g., "What do you think is the most important problem facing this country today and which political party can solve this best?"). TTB merely makes use of the "best" available piece of knowledge in a specific circumstance. TTB operates around two guiding ideas. According to the first recognition principle, if only one among a number of alternatives is recognized when making a decision under ambiguity, then the recognized alternative should be selected [3]. In other words, if there is only one rider you are familiar with in the race, pick that horse. When more than one alternative is identified and the recognition principle is unable to give discriminatory information, the second principle is applied.

IV. RESEARCH DESIGN

How does machine learning help us predict the outcome of elections? Machine learning has traditionally been used to make predictions about other things, such as the stock market or fraud patterns. How is it different from traditional predictive modelling?

Traditional predictive modelling [6] relies on probabilistic formulas and statistical analysis. In contrast, our method can be implemented by programmers, and requires no prior knowledge of statistical analysis. We are also able to create useful forecasts and provide insights for the prediction of elections based on the variables we analysed, without any prior knowledge of what the outcome might be. In the future, how might machine learning help us understand political events? Prediction is a good way to understand political trends as discussed by [7]. With machine learning, prediction models can be built to assess any factor of who and how people vote. This combined with big data analytics can create a model predicting the outcome of elections, giving us a better perspective of the future. Machine learning algorithms can help in predicting the outcomes of elections with more accuracy than ever before. It takes in all the data available and analyses it systematically using the set of variables that should be defined previously as impacting factors. The identified impacting factors are: party, Ethnicity, Age, Area, Net worth, marital status, income, city.

In theory, as discussed by [5,8] learning algorithms can help improve analysis results and generate robust statistical models that can be used to predict other types of data, such as stock prices. In practice, it is difficult to obtain a suitable model to perform useful predictive analysis for a specific situation. In order to reduce this problem, statistical models are usually used to describe the statistical process called latent logic. In this model, the voter's movement is represented by latent variables.

These variables [5] are related to the voter's movement, behaviour, and response, but are not part of the voter itself. This method is very popular when modelling geopolitical situations with complex dynamics in the system. This is the method we use to understand the dynamics of voting behaviour. The analyses on impacting factors on election prediction is further analysed, and insights and results have been provided. The devised method pinpoints whether different ways of collecting different data of election voters can lead to much better prediction and understanding of the election process [6]. Therefore, we consider above mentioned categories as a relevant factor to be measured for prediction of elections.

The results show that when the survey measure the above criteria's defined as impacting factors, they prefer to build and keep key connections to deliver vital information and enhance the likeliness of prediction of the election outcome. This means that researchers who would like to predict elections should collect the correct information prior to process it further and collect the data and therefore to plan an activity of prediction and knowledgably participate in forecasting process [7]. Based on the information presented above, the goal is for them to be able to decide who to add as a category. Having strategic categories identified as impacting factors can also improve the results of the forecasting the proper results and outcome of the elections.

V. DATA ANALYSES

In theory, learning algorithms can help improve analysis results and generate robust statistical models that can be used to predict other types of data, such as stock prices.

In practice, it is difficult to obtain a suitable model to perform useful predictive analysis for a specific situation. In order to reduce this problem, statistical models are usually used to describe the statistical process called latent logic.

In this model, provided in the table 1 below, the voter's movement is represented by latent variables. These variables are related to the voter's movement, behavior, and response, but are not part of the voter itself.

TABLE 1. POSTERIOR DISTRIBUTION CHARACTERIZATION FOR PAIRWISE CORRELATIONS[A]

			prob_winner_is	age
prob_winner_is	Posterior	Mode		,059
		Mean		,058
		Variance		,005
	95% Credible Interval	Lower Bound		-,079)
		Upper Bound		,194
	N		200	200
age	Posterior	Mode	,059	
		Mean	,058	
		Variance	,005	
	95% Credible Interval	Lower Bound	-,079)	
		Upper Bound	,194	
	N		200	

a. The analyses assume reference priors (c = 0).

In the Table 1. provided is the bivariate normal distribution-based Pearson correlation coefficient assesses the linear relationship between two scale variables “age” and “probable_winer_is”, all together. Here, we concentrate on a Bayesian hierarchical model that enables to deduce the underlying association between observations that have been tainted by errors. We demonstrate that this method can also be used to determine the underlying connection between uncertain parameter estimates and the correlation between uncertain parameter estimates and noisy data. With a collection of empirical data, we demonstrate the Bayesian modeling of correlations. To quantify the proof that the data support the existence of a relationship, we first estimate the posterior distribution of the underlying correlation and then compute Bayes factors.

TABLE 2. BINOMIAL PROPORTION

	Posterior			95% Credible Interval	
	Mode	Mean	Var.	Lower Bound	Upper Bound
party	,144	,147	,001	,102	,199
ethnicity	,188	,191	,001	,140	,248
age	,010	,015	,000	,003	,035
city	,030	,034	,000	,014	,063
area	,554	,554	,001	,485	,621
Net worth	,505	,505	,001	,436	,573
income	,342	,343	,001	,280	,410
Marital status	,050	,054	,000	,027	,089

a. Prior on Binomial proportion: Beta(2, 2).

In the above Table 2 provided are all the main variables we have measured. We have provided posterior binomial proportion This method is very popular when modelling geopolitical situations with complex dynamics in the system. This is the method we use to understand the dynamics of voting behavior. In order to do the analyses we have used the SPSS Statistics software package used for statistical analysis.

TABLE 3. DESCRIPTIVE STATISTICS OF WITHIN-SUBJECT FACTOR LEVELS

Dependent Variables	Mean	Std. Deviation	N	Min	Max
prob_winner_is	2,36	1,093	109	1	4
age	57,12	17,052	109	21	91

In the above Table 3 provided are descriptive statistics of within-subject factor levels for assessment of the linear relationship between two scale variables “age” and “probable_winer_is”.

TABLE 4. BAYESIAN ESTIMATES OF GROUP MEANS[A]					
Dependent Variables	Posterior			95% Credible Interval	
	Mode	Mean	Variance	Lower Bound	Upper Bound
Prob winner is	2,36	2,36	1,327	0,10	4,62
age	57,12	57,12	1,327	54,86	59,38

a. Posterior distribution was estimated based on the Bayesian Central Limit Theorem.

In the above Table 4 provided are the bayesian estimates of group means descriptive statistics of within-subject factor levels for assessment of the linear relationship between two scale variables “age” and “probable_winer_is”.

The Posterior distribution was estimated based on the Bayesian Central Limit Theorem. Although some of the components are shared by all of the proofs of the finite-dimensional Bayesian central limit theorem, our approach is mostly based on the demonstration of Theorem 1.4.2 in Ghosh and Ramamoorthi's book Bayesian Nonparametrics (BCLT, aka the Bernstein-von Mises theorem) discussed by [10].

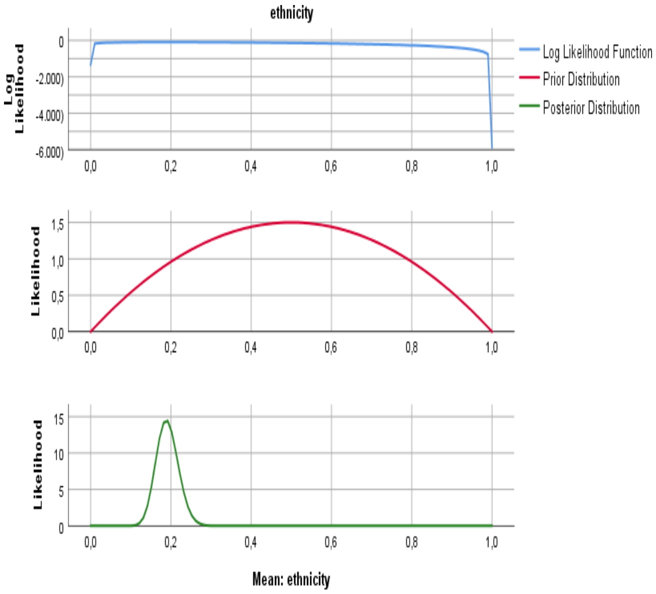

Fig. 1. Posterior distribution by ethnicity

In the figure 1 the Posterior Distribution of Group Means chart helps visualize the difference between the posterior distributions. Posterior distribution by ethnicity is the focus, and it may be worthwhile to further investigate and compare additional variables in future studies.

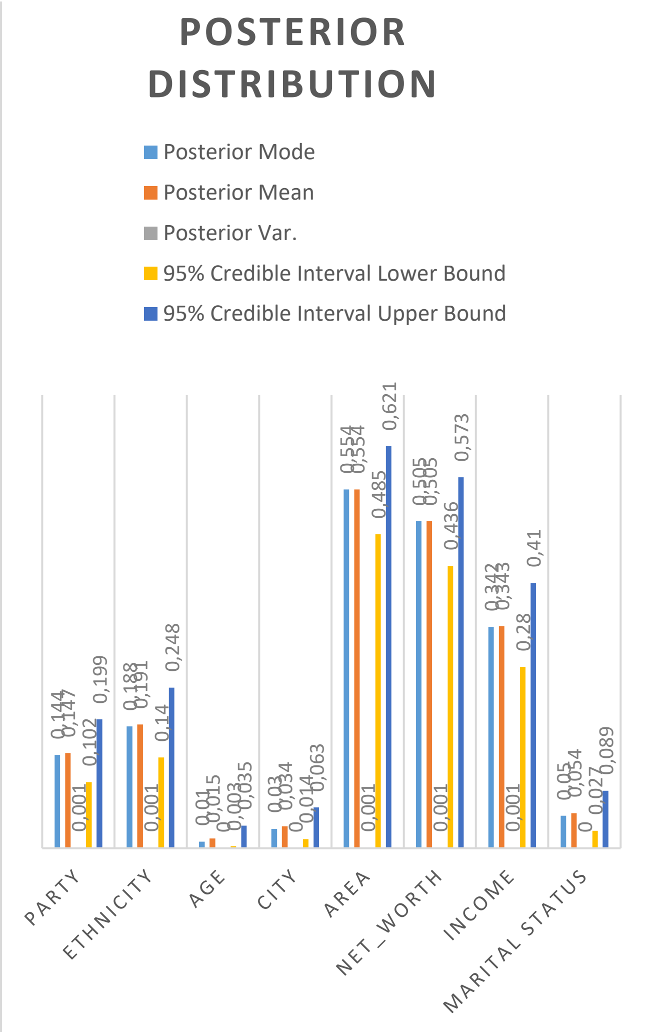

Fig. 2. Posterior distribution by Posterior Mean and Credible interval

In the figure 2 provided is the posterior distribution by 8 categories: party, ethnicity, age, city, area, Net worth, income, Marital status. The graphic calculates the 95% credible interval for lower bound and upper bound. The alpha value for a variable “Prob winner is” is 0,10, and the associated critical value is 4,62for a two-tailed 95% confidence interval. Also, alpha value for “age” is 54,86 and the associated critical value is 59,38 This indicates that we can subtract 2.36 standard deviations from the mean, or 57,12 for the mean of age, to determine the upper and lower boundaries of the confidence interval.

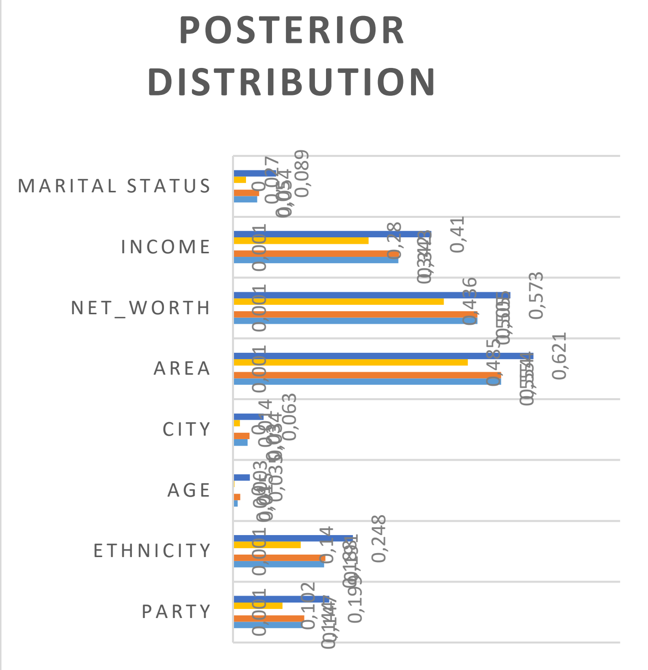

Fig. 3. Posterior distribution by 8 categories

In the figure 3 above provided is the posterior distribution fro the 8 (eight) categories. The posterior probability distribution often expresses the epistemic uncertainty regarding statistical parameters conditional on a set of observed data in the context of Bayesian statistics. Consider a party with a voter base that is 60% men and 40% women. Equal numbers of girls and boys dress in skirts or pants. The only thing a far observer can make out about a (random) voter is that they are wearing pants. What is the likelihood that this voter is a female? The Bayes theorem can be used to calculate the right response.

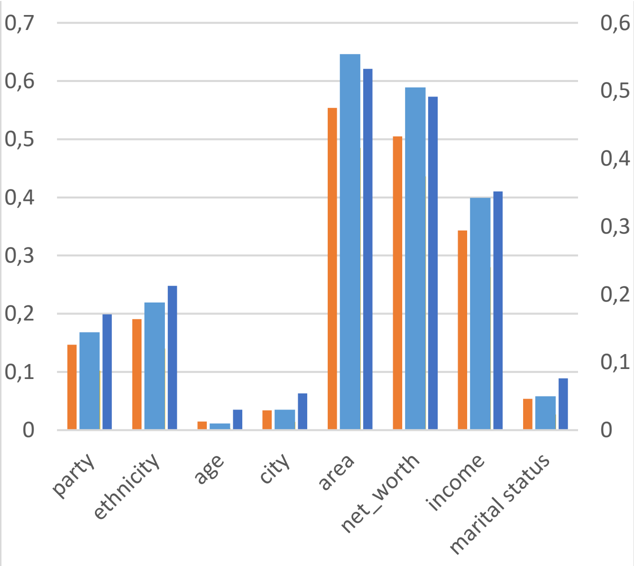

Fig. 4. Posterior distribution averages

In the figure 4 above provided is the Posterior distribution averages for the measured variables from the study. The posterior probability, which is taken into consideration by posterior distribution averages, is the likelihood that an event will occur after all available data and contextual information have been considered. It is closely related to prior probability, which is the likelihood that an event will occur before any new evidence has been considered. The posterior probability can be viewed as a correction to the prior probability: For instance, according to previous research, 60% of voters who enroll in college will graduate within six years. The prior probability is as stated. You believe that number to be significantly lower, so you set out to get additional information. The posterior probability is that the true number is actually closer to 50% based on the facts you have gathered.

VI. Discussion of the Results

The Bayesian Estimates of Group Means table summarizes the posterior statistics based on the findings for each repeated-measure variable, including the 95% credible intervals. The output includes basic descriptive statistics of the data set and the repeated-measure variables that are specified for the procedure.

With the use of a group variable, it is possible to identify two unrelated groups and draw conclusions about their differences using the Bayesian Independent - Sample Inference process. You can characterize the desired posterior distribution either by assuming the variances are known or unknown and estimate the Bayes factors using a variety of methods.

VII. Conclusions

We have developed a method for recognizing the main impacting factors in elections by using Survey dataset that was originally collected within the study. The data are collected from Surveys because politicians and mainly voters, do not use substantially Twitter, Facebook or Instagram and most of the analytics is based on input from Survey. In order to devise and asses the impacting factors we have devised most-important-problem (MIP) algorithmic model to predict the outcome of upcoming future Kosovo parliamentary elections and grounded it on the TTB (take-the-best) strategy.

An analysis of forecasting approach to elections and the performance metrics (variance) using the (MIP) algorithmic model has been used. The data has derived from an originally collected survey dataset that contains the impacting factors previously identified and assessed regarding the previous parliamentary elections in Kosovo 2021 has been realized.

Discussion of results for each particular statistical analyses has been realized in the previous Data Analyses section.

References

[1] Czerlinski, J., Gigerenzer, G. & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd & the 272

[2] ABC Research Group (Eds.), Simple heuristics that make us smart (pp. 97-118). Oxford University Press. 273

[3] Gelman, A., Goel, S., Rivers, D., & Rothschild, D. (2016). The mythical swing voter. Quarterly Journal of Political Science, 11(1), 274, 103–130. 275

[4] Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological 276 Review, 103, 650– 669 277

[5] Graefe, A. (2015). Improving forecasts using equally weighted predictors. Journal of Business Research, 68(8), 1792–1799. 278

[6] Graefe, A., Armstrong, J. S., Jones, R. J. J., & Cuzán, A. G. (2014). Combining forecasts: An application to elections. International 279, Journal of Forecasting, 30(1), 43–54. 280

[7] Graefe, A., Armstrong, J. S., Jones, R. J. J., & Cuzán, A. G. (2017). Assessing the 2016 U.S. presidential election popular vote 281

[8] Ana Elsa Hinojosa Herrera, Chris Walshaw and Chris Bailey, (2020) Failure Mode & Effect Analysis and another Methodology for Improving Data Veracity and Validity, Annals of Emerging Technologies in Computing (AETiC), Vol. 4, No. 3, 2020