Predictive Analysis of United States Presidential Elections Using K-Prototype Clustering Honors Capstone

Sebastian Munoz and Edgar Franco-Vivanco[1]

Department of Industrial and Operations Engineering, College of Engineering, University of Michigan

INTRODUCTION

The primary objective of this project is to create a predictive modeling tool that can be used in order to forecast future United States presidential elections based on a clustering algorithm.

Some of the research questions that this project seeks to explore and answer are listed below:

- Is it possible to predict which way a state will vote?
- What are some key factors that affect election results?
- What population indicators separate clustering groups?

METHODS

Based on partitioning, the K-Prototype clustering algorithm is an improvement of the K-Means and K-Mode clustering models by handling mixed data types. In other words, it is capable of analyzing both numerical and categorical features to arrive at a more thorough classification of clusters.

Step 1	Read in feature selected key population indicators, comprised of both categorical and numerical data variables
Step 2	Initialize prototype selection by selecting k points as the initial prototypes for k clusters at random, considering the dissimilarity measure
Step 3	Initialize clustering allocation by assigning each object to a cluster which has the minimum difference with its prototype with the previous method
Step 4	Iterate through the algorithm and reallocate data points until the different moves is unchanged, indicating a best result

The dataset, which was extracted from United States government databases, included detailed 2020 census data, historical election counts for every election from 1976-2020, and historical election results for every state from 1976-2020.

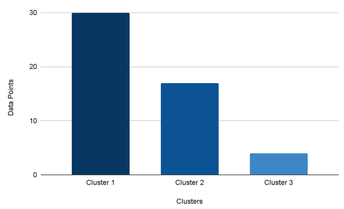
The Elbow Method was then utilized in order to determine the optimal number of clusters that should be implemented. This is reflected by the K value corresponding to the point where the graph starts to move almost parallel to the X-axis. The initial results found the optimal number of clusters to be K = 3.

From this, a weighted forecasting model was then adopted in order to predict the 2024 United States presidential election. In addition to this, a comprehensive analysis of the associated parameters was then performed in order to better understand the clustering results and identify key population indicators.

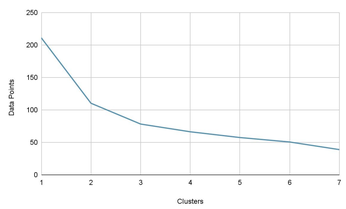
RESULTS

The finalized clustering results are depicted below:

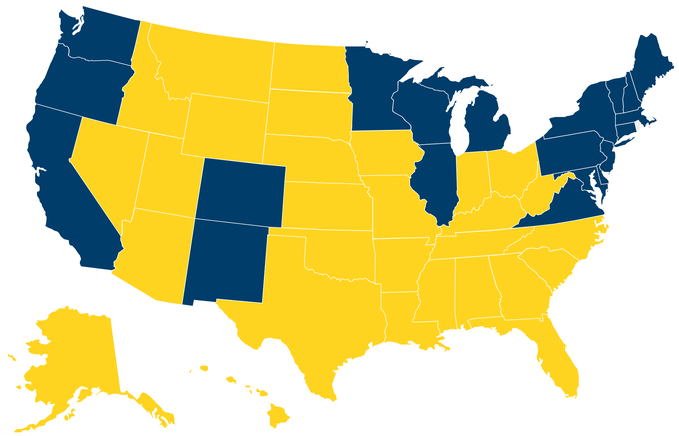
Final Clustering Results



Elbow Method



The subsequent forecasting prediction for the 2024 United States presidential election, broken down by state, is as follows:



REPUBLICAN	DEMOCRAT
276	262

The numbers shown above reflect the final predicted outcome of the Electoral College process, which consists of 538 electors divided among all 50 states and the District of Columbia based on the same number of electors as they have in its Congressional delegation. A majority of 270 votes is required to win the election. Therefore, it is possible to conclude that the Republican Party will win the 2024 United States presidential election.

ASSESSMENT

Although this project was able to successfully develop a predictive modeling tool that can be used in order to forecast future US presidential elections, it is also important to understand and analyze the various characteristics that comprise the different clustering groups. That being said, a detailed breakdown of several key factors is provided within the following section.

ASSESSMENT

The results of the cluster-based, weighted forecasting model for the 2024 United States presidential election is summarized in the table below. It is interesting to note that cluster 1 constitutes the majority of Republican-voting states whereas cluster 2 represents the majority of Democrat-voting states.

Party	Cluster 1	Cluster 2	Cluster 3	Total
Democrat	7	11	2	20
Republican	23	6	2	31
Total	30	17	4	51

These results can be attributed to the clustering algorithm which determines a states allocation by considering the dissimilarity measure for each associated variable. This project then began to asses the key population indicators that differentiate clusters, with some of the summarized results shown below:

Various analysis methods were then completed across several relevant population indicators, granting many unique insights. Some of the interesting variable deep-dives include the native born population rate, bilingual rate, poverty rate, median household income, occupational breakdown, health insurance coverage breakdown, and cluster industry breakdown.

Cluster 1	• Lowest bilingual rate • Largest native born population • Smallest average population size • Most households without internet access
Cluster 2	• Highest median household income • Largest rates of educational attainment • Lowest jobless rate • Smallest poverty rate
Cluster 3	• Highest bilingual rate • Smallest native born population • Largest average population size • Youngest median population age

This analysis helps in understanding the forecasted presidential election results and confirms the validity of the K-Prototype clustering-based predictive modeling tool. For additional information, please contact sebmunoz@umich.edu.

ACKNOWLEDGEMENTS

Edgar Franco-Vivanco[1]
Assistant Professor of Political Science
Rachel Armstrong-Ceron
Senior Academic Advisor