Group 4 Final Project

1st Riley Hawley
Dept. of Computer Science
Stevens Institute of Technology
Hoboken, United States
rhawley2@stevens.edu

2nd Eddie Kuang
Dept. of Computer Science
Stevens Institute of Technology
Hoboken, United States
ekuang@stevens.edu

3rd Jince Shi
Dept. of Computer Science
Stevens Institute of Technology
Hoboken, United States
jshi39@stevens.edu

4rd Eric Tashji
Dept. of Computer Science
Stevens Institute of Technology
Hoboken, United States
etashji@stevens.edu

***Abstract*—The problem statement for the project is: "Republican voters tend to reside more in rural areas, whereas Democrat voters tend to reside more in urban areas. Republicans want to increase its turnout in urban areas by 5% and Democrats want to increase its turnout in rural areas by 5%." The Machine Learning algorithms we are using to solve the problem are the K-means, Logistic Regression, Linear Regression, and Decision Tree algorithms. Thus far, our experimental results indicate that while rural voters tend to vote Republican and urban voters tend to vote Democrat, a big factor that plays into that result is the variance in the individuals' most important issues.**

I. Introduction

The problem statement for the project is: "Republican voters tend to reside more in rural areas, whereas Democrat voters tend to reside more in urban areas. Republicans want to increase its turnout in urban areas by 5% and Democrats want to increase its turnout in rural areas by 5%."

There will be two data sets: one for voters in rural areas and one for voters in urban areas. The data will include each individuals age, gender, marital status, number of children, occupation, salary, standard of living, party membership, rate of voter turnout, most important issue (when polled), engagement with Democrat or Republican ground campaign, and percentage of votes for Democrat or Republican candidates for the past 8 years.

We will use four machine-learning algorithms to solve the problem: the K-means algorithm, the logistic regression algorithm, the linear regression algorithm, and the decision tree algorithm. The first three algorithms will be used to find significant data trends, and the final algorithm will be used to figure out what decisions could be made to increase voter turnout for each respective side (Democrat or Republican).

Thus far, our experimental results indicate that big headways could be made on both sides by two factors. The first is to increase those parties' respective ground involvement with voters of the opposite party, and the second factor is to increasingly address the issues that are most important to a voter. As far as how to target with the ground game with regards to voter turnout, the results are inconclusive. While those with a lower voter turnout may be more receptive to a change in opinion, they are also less likely to actually go out and vote.

Many of the existing solutions involve enhancing engagement, and providing various forms of civic education. Civic education; especially, is useful in changing voters' minds. This is because in allowing voters to understand how the political systems work, it alleviates their fears surrounding the opposite parties' policy positions, as well as ensuring they don't think they are giving power to people they believe should not have it.

II. Related Work

Solutions to this problem can be organized into two different categories: increasing engagement, and increasing civic education.

Solutions for increasing engagement include increasing communication, giving surveys, providing voting incentives, and maintaining contact after the election. Frequent engagement doesn't just encourage voter turnout, but also provides voters with the opportunity to engage with representatives of the other party, who may be able to answer their questions and allow them to see things from a different perspective.

Solutions for increasing civic education include civic education campaigns, as well as clear explanations of the positions of party candidates. This helps to ease voters' fears regarding the policies of various candidates and allows them to consider different perspectives. Furthermore, increasing civic education allows voters to feel more secure in their vote as they will understand how the checks and balances system works.

III. Our Solution

A. Description of Dataset

The datasets used in this analysis were derived from voter turnout and population density data across various counties. The voter turnout data comes from the official primary election held on June 4, 2024, while the population density data was sourced from Census 2020 and 2022 estimates.

During preprocessing, a few issues in the dataset were identified and corrected. Some records had missing values in the "Number of Children" field, which could have affected analyses related to family size and voting behavior. Additionally, there were extreme values in the "Salary" column that could have distorted our findings.

To address these issues, missing values in the "Number of Children" field were filled using the median, ensuring the data remained representative. Outliers in the "Salary" column were capped at a reasonable level to prevent skewed analysis.

Categorical data, such as gender, marital status, occupation, and party membership, were converted into numerical values

for consistency. Continuous features like age and salary were normalized to ensure each feature contributed equally to the analysis. Data visualization techniques were used to check distributions, ensuring the data was ready for accurate analysis. These preprocessing steps cleaned and standardized the dataset, allowing for meaningful insights into voting patterns and demographics.

B. Machine Learning Algorithms

Here, we will implement four machine learning algorithms. We will implement K-mean clustering, logistic regression, linear regression and decision tree to gain insights of the data. K-means will be used to divide the data into clusters to look for some trends and patterns into urban and rural voters. K-means is an appropriate to look for patterns and insights on data. Either the Silhouette score or the Elbow method will be used to determine to number of clusters or both to compare the difference. Logistic regression and decision tree will be used to predict if people in rural will turnout and also to be used to see if urban people will be predicted to be turnout or not and use that result to see if we can increase the turnout for both groups by 5%. We will use gridsearchCV or similar to find the optimal hyperparameters for the algorithms. These classification algorithms are appropriate because, they will predict to see if the people will likely turnout or not and ultimately predict if we can increase turnout for both groups by 5%. Linear regression can be use to to explore the actions that will increase turn out rates. This can be used to see if the turnout rate is increased as other parameters increases.

1) Linear Regression: We plan to employ linear regression to model the relationship between various demographic factors and voter turnout rate. This method is chosen because:

- It will allow us to quantify the impact of multiple independent variables on voter turnout.
- The coefficients will provide interpretable insights into the importance of each factor.
- It can handle both continuous and categorical variables (after appropriate encoding).

For our analysis, we will focus on key numerical variables including age, salary, number of children, and voter turnout rate. We plan to normalize these features to ensure they are on the same scale. Categorical variables like gender, marital status, and occupation will be one-hot encoded to be usable in the linear regression model.

In our preliminary analysis, we observed that the age distribution differs between urban and rural areas, which may impact voting patterns. We are currently working on visualizations to illustrate these differences.

Our linear regression model will take the form:

$$\text{Voter Turnout Rate} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Salary}) + \beta_3(\text{NumChildren}) + \cdots + \epsilon \tag{1}$$

Where β_0 is the intercept, β_1, β_2, etc. are the coefficients for each feature, and ϵ is the error term.

2) Logistic Regression: We plan to use logistic regression to help us with two main tasks:

- Predicting Voter Turnout: Well use it to figure out if a voter is likely to show up at the polls based on different demographic and engagement factors.
- Predicting Party Affiliation: It will also help us predict whether a voter is more likely to vote Republican or Democrat using the same set of factors.

We chose logistic regression for these tasks because:

- Its great for binary classification problems, which is perfect for our yes/no questions about voter turnout and party choice.
- The results are easy to understand, as the coefficients show how different factors influence the likelihood of turnout and party preference.
- It can work with both numerical data (like age and salary) and categorical data (like marital status and party membership) after we encode the categories properly.

For our analysis, we'll focus on important variables such as age, salary, number of children, marital status, occupation, party membership, and engagement with ground campaigns. We'll convert categorical variables into numerical values (one-hot encoding) and normalize continuous features like age and salary to make sure everything is on the same scale.

C. Implementation Details

1) Linear Regression Implementation: We are in the process of implementing the linear regression model using scikit-learn in Python. Our planned process involves:

1) Preprocessing the data, including handling missing values and encoding categorical variables.
2) Splitting the data into training (80%) and testing (20%) sets.
3) Fitting the model on the training data.
4) Evaluating the model's performance on the test data using R-squared and Mean Squared Error (MSE) metrics.

While we haven't yet obtained results, we anticipate that our analysis will reveal differences in the factors influencing voter turnout in urban versus rural areas. We expect variables such as age, education level, and income to have significant impacts, but the magnitude of these effects may vary between urban and rural settings.

In our next steps, we plan to:

- Complete the data preprocessing steps
- Implement the linear regression model
- Analyze the results and create visualizations to illustrate our findings
- Compare the performance of linear regression with other machine learning models our team is implementing

So far, for fitting a basic k-means and using the Silhouette score, it found optimal k to be 3, so 3 different distinct groups. I'm thinking of also trying GMM alongside of the k-means because of the nature of k-means having a "hard" answer of having to belong in a group as oppose to GMM with a more probablistic approach.

IV. Comparison

This section includes the following: 1) comparing the performance of different machine learning algorithms that you used, and 2) comparing the performance of your algorithms with existing solutions if any. Please provide insights to reason about why this algorithm is better/worse than another one.

V. Future Directions

This section lays out some potential directions for further improving the performance. You can image what you may do if you were given extra 3-6 months.

VI. Conclusion

This section summarizes this project, i.e., by the extensive experiments and analysis, do you think the problem is solved well? which algorithm(s) might be better suitable for this problem? Which technique(s) may help further improve the performance?

Last but not the least, don't forget to include references to any work you mentioned in the report.