A Tale of Two Parties in Urban and Rural America

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 Gaussian Mixture Model, 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 individual's 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 Gaussian Mixture Model, 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 Gaussian Mixture Model clustering, logistic regression, linear regression and decision tree to gain insights on the data. 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:

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

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: Logistic regression was used to tackle two main tasks:

- Predicting Voter Turnout: This model helped determine if a voter was likely to show up at the polls based on various demographic and engagement factors.
- Predicting Party Affiliation: It also assisted in predicting whether a voter was more likely to vote Republican or Democrat using the same set of factors.

Logistic regression was chosen for these tasks because:

- It is well-suited for binary classification problems, which made it ideal for our yes/no questions about voter turnout and party affiliation.
- The results were easy to interpret, as the model's coefficients indicated how different factors influenced the likelihood of turnout and party preference.
- It was compatible with both numerical data (like age and salary) and categorical data (like marital status and party membership) after appropriate encoding.

In this analysis, the focus was on key variables such as age, salary, number of children, marital status, occupation, party membership, and engagement with ground campaigns. Categorical variables were converted into numerical values through one-hot encoding, and continuous features like age and salary were normalized to ensure consistent scaling across all features.

3) Gaussian Mixture Model: Initially, for clustering and finding patterns in the data, the k-means algorithm was going to be used, however, Gaussian Mixture Model was used instead because of it's probabilistic nature assigns a probability to the data instead of a hard decision of having to belong in a group and because it allows for elliptical shape instead of just spherical. The clustering algorithm GMM, is good for identifying some patterns in the data of voters in both urban and rural, especially with so many data points. It can help us identify which features make up a certain cluster groups.

4) Decision Tree: The decision tree was chosen to determine the following:

- Determining the ratio of Republican to Democrat in rural and urban areas.
- Determining the rate of voter turnout within each party.

Decision Tree was chosen for these tasks because:

- A decision-tree can be used to breakdown the distribution of party affiliation in urban and rural areas.
- A decision-tree can be used to breakdown the distribution of likely/unlikely voters.
- A decision-tree can be used to determine the breakdown of likely/unlikely voting within each party.

In building the decision-tree, the results were focused on determing the factors associated with an individual's party affiliation.

C. Implementation Details

1) Linear Regression Implementation: We implemented three linear regression models to analyze voter turnout and party preference in urban and rural areas. The implementation process involved several key steps:

1) Data Preprocessing:

- Handled missing values in the 'Number of Children' column by filling with the median.
- Converted categorical 'Rate of Voter Turnout' to numerical values (Low: 0, Medium: 1, High: 2).
- Created a PartyPreferenceScore for party preference analysis.

2) Feature Selection:
   - For voter turnout: Used both numeric (Age, Number of Children, Salary) and categorical features (Gender, Marital Status, Occupation, Standard of Living, Party Membership, Most Important Issue, Engagement with Campaign).
   - For party preference: Initially used only numeric features, then expanded to include categorical features.
3) Model Training:
   - Split data into 80% training and 20% testing sets.
   - Used StandardScaler for numeric features and OneHotEncoder for categorical features.
   - Employed sklearn's Pipeline and ColumnTransformer for preprocessing.
4) Model Evaluation:
   - Used Mean Squared Error (MSE) and R-squared (R2) score for performance evaluation.
   - Performed cross-validation for the voter turnout model.

Our analysis yielded the following results:

1) Voter Turnout Model:
   - Urban: R2 Score: 0.3116, Mean CV R2 score: 0.3081
   - Rural: R2 Score: 0.3214, Mean CV R2 score: 0.3081

   Top features for Urban areas:
   - Occupation (Real Estate Agent, Software Developer)
   - Most Important Issue (Education)
   - Standard of Living (Low)

   Top features for Rural areas:
   - Party Membership (Republican)
   - Most Important Issue (Healthcare)
   - Occupation (Healthcare Worker, Teacher)
2) Party Preference Model (Limited Features):
   - Urban: R2 Score: 0.0007
   - Rural: R2 Score: 0.0009

   Top feature for both areas: Rate of Voter Turnout
3) Party Preference Model (Extended Features):
   - Urban: R2 Score: 0.7291
   - Rural: R2 Score: 0.7563

   Top features for both areas:
   - Party Membership (Republican)
   - Occupation (various)
   - Marital Status (Single, for rural areas)

The voter turnout model shows moderate predictive power, explaining about 31-32% of the variance in both urban and rural areas. This suggests that while our selected features do influence voter turnout, there are likely other important factors not captured in our data.

The limited-feature party preference model performed poorly, indicating that Age, Number of Children, Salary, and Rate of Voter Turnout alone are not good predictors of party preference.

The extended-feature party preference model showed significantly better performance, explaining about 73-76% of the variance in party preference. This highlights the importance of including categorical variables, especially Party Membership, in predicting voting behavior.

In both urban and rural areas, Party Membership (Republican) was the strongest predictor of party preference, which is expected but also indicates strong party loyalty. Occupational factors play a significant role in both voter turnout and party preference, with different occupations being influential in urban versus rural settings.

The average party preference scores (Urban: 0.2088, Rural: -0.0836) confirm the tendency for urban areas to lean Democratic and rural areas to lean Republican.

Our linear regression analysis reveals that while demographic factors do influence voter turnout and party preference, party membership and occupation are particularly strong predictors. The models perform better at predicting party preference than voter turnout, suggesting that turning out to vote is a more complex behavior to predict than party allegiance. To increase turnout, parties might focus on occupation-specific outreach and addressing key issues like education in urban areas and healthcare in rural areas. For increasing appeal in opposition areas, parties should consider tailoring their message to the occupations that show stronger alignment with the opposing party in those areas.

2) Logistic Regression Implementation: Logistic regression was implemented to tackle the tasks of predicting voter turnout and party affiliation in both rural and urban areas. The process involved several key steps to ensure that the model was both effective and interpretable.

1) Data Preprocessing: The initial phase of implementation focused on addressing data quality issues. For instance, missing values in the "Number of Children" feature were imputed using the median, which was appropriate given the skewness of the distribution. Additionally, extreme outliers in the "Salary" feature were capped to prevent distortion of the model's predictions.

   Categorical variables such as gender, marital status, occupation, and party membership were transformed into numerical formats through one-hot encoding. This step was crucial for logistic regression, which requires numerical inputs. Continuous features, including age and salary, were normalized to ensure that they contributed equally to the model's predictions.

2) Model Training and Testing: The dataset was split into an 80% training set and a 20% testing set

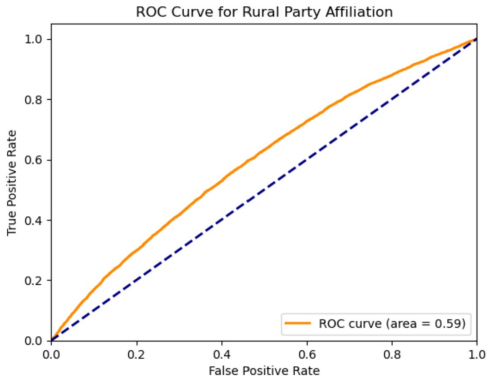

Fig. 1. ROC Curve for Rural Party Affiliation

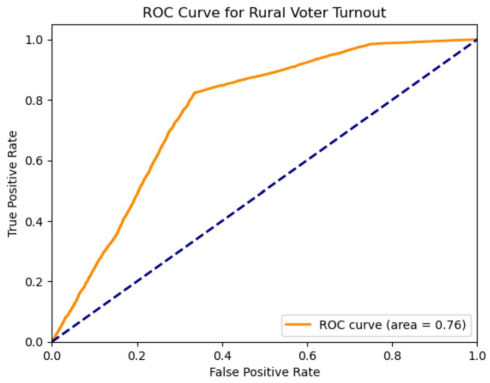

Fig. 2. ROC Curve for Rural Voter Turnout

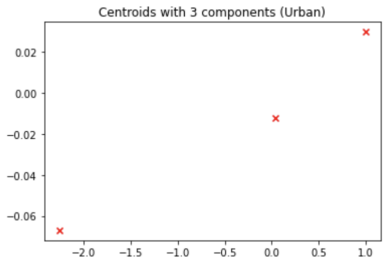

Fig. 3. Centroid Cluster for Urban-3

to evaluate the model's generalization ability. The model was built using the SGDClassifier with a logistic regression loss function, which is known for its efficiency in handling large datasets.

The model was trained on the training set and its performance was evaluated on the testing set. The evaluation provided insights into the model's accuracy and its ability to predict voter turnout and party affiliation.

3) Evaluation Metrics: The performance of the logistic regression model was assessed using metrics such as precision, recall, F1-score, and accuracy. These metrics offered a comprehensive evaluation of the model's classification performance.

For rural data, the model achieved an accuracy of 59% for predicting party affiliation, with a precision of 0.60 for predicting Republican affiliation and 0.55 for predicting Democrat affiliation. The voter turnout model performed better, with an accuracy of 76%. The ROC AUC scores were 0.59 for party affiliation and 0.76 for voter turnout, indicating that the model was more reliable in predicting voter turnout than party affiliation.

In urban data, the model's accuracy for predicting party affiliation was 69%, with a stronger performance in predicting Democrat affiliation (precision of 0.70) compared to Republican affiliation (precision of 0.51). The voter turnout model had an accuracy of 77%, with a ROC AUC of 0.61 for party affiliation and 0.75 for voter turnout.

4) Interpreting Coefficients: One of the advantages of logistic regression is the interpretability of its coefficients. For rural data, the coefficients revealed that certain occupations like retail workers were positively associated with Democrat affiliation, while occupations such as police officers were more aligned with Republican affiliation. Similarly, in urban data, high-salary occupations like finance advisors and healthcare workers were associated with Democrat affiliation, while IT specialists and software developers leaned towards Republican affiliation.

In terms of voter turnout, higher salaries were negatively associated with turnout in both rural and urban areas, while engagement with the campaign had a positive influence on voter turnout. These insights are valuable for understanding the demographic factors that influence voter behavior.

3) Gaussian Mixture Model: For the clustering, Gaussian mixture model was used on the data to identify some characteristics about the data. There were 2 separate datasets so each dataset was perform using the following blueprint. First, the data had missing values and checking the skewness of the features, a lot of them had some skewness. Because of the skewness, the median was used to impute the missing values. All the categorical values were converted to integers. Eg. 0 for Single, 1 for married, 2 for never married etc. **It is important to keep in mind here, that party affiliation for this model was converted so, 0 represents republican and 1 represents democrat.** Each feature was standardized as well before performing anything. There was significant amount of features so principal component analysis was used to reduce the number of components. A plot of the cumulative sums on the explained variance ratio was used to determine the number of components. We want the number of components to explain between 90%-95%. AIC was used to initially look at the best number of clusters for the Gaussian mixture model. Plotting the centroid was used to see if the clusters are well separated. The mean values of each feature was used to group each cluster to give a general overview of how each cluster was split.

For the urban dataset, the Number of components vs the

Cluster	Age	Gender	Marital Status	Number of Children	Occupation
0	53.987892	0.574418	1.501406	2.470703	9.431493
1	54.048862	0.576017	1.498727	2.473116	9.641882
2	54.303936	0.576438	1.483829	2.477594	9.482920

Cluster	Salary	Standard of Living	Party Membership
0	49961.077672	1.001485	1.000000
1	49994.875508	0.999745	0.000000
2	502959.490814	0.994415	0.702819

Cluster	Rate of Voter Turnout	Most Important Issue
0	1.080062	2.001836
1	1.090456	2.002413
2	1.083907	2.017665

Cluster	Engagement with Campaign	Votes for Democrat (Last 8 years)
0	1.100196	6.003769
1	1.199827	2.003160
2	1.124692	4.791401

Cluster	Votes for Republican (Last 8 years)
0	1.998375
1	5.999474
2	3.175347

Cluster	Votes for Other Candidate (Last 8 years)
0	0.999309
1	1.000200
2	0.997272

Fig. 4. Mean of Urban Clusters Results

variance explained plot suggest PCA should use 10 components for achieving 90%-95% of the variance explained. The AIC for the urban data suggest that 6 clusters was the best. After fitting 6 clusters for the Gaussian mixture model, there was a very noticeable overlap between clusters, which might suggest overfitting even though AIC accounts for overfitting. 4 clusters were used next to fit another GMM model and there were still 2 clusters very close to each other so further reduction of clusters was done. The AIC between 3 and 4 was very minimal which suggest 3 will fit better, especially since there is still overlap with 4. After fitting 3 clusters, the plot of centroids looks a lot better with each clusters separated pretty well.

After putting the clusters with the original data, we can notice a few things. Age, gender, Marital status, number of children, standard of living, rate of voter turnout, most important issues, and votes for other candidates last 8 years didn't seem to be significant in determining cluster assignments (there wasn't much difference in the averages). Occupation, engagement with campaign had some significances in determining cluster assignments with some differences in the mean. The features that had the most significances in determining cluster assignments was salary, party membership, votes democrat within last 8 years, votes republicans for last 8 years with big differences in the mean. Another noticeable difference in the cluster was a cluster with a mean salary of over $500,000. Looking into the original data, for urban, there were 7370 people earning salaries greater than $250,000. Based on the cluster plot, there were 2 very noticeable clusters and 1 not very noticeable cluster. It is hard to see in the plot but it seems one of the clusters is the group with a very high salary but it is hard to see because it is only 7370 out of over 700000 people.

It seems the clusters of voters for urban were mainly determined by salary, party membership, votes democrat within last 8 years, votes republicans for last 8 years. The first cluster had relatively lower salary compared to cluster 2 but similar to cluster 1. Party member had significant differences between cluster 0 and cluster 1. Cluster 0 is mainly democratic and cluster 1 is mainly republicans. We can see the higher salary earners tend to lean towards being democrats. The mean for cluster 2 for party member is 0.7, which tends to lean towards cluster 2 being more democrats. Looking at votes for democrat in the last 8 years, we can see cluster 0 the mean for voting

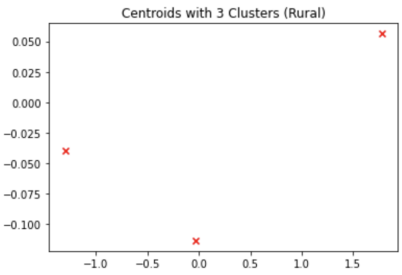

Fig. 5. Centroid Cluster for Rural-3

democrats in the last 8 years was 6. For cluster 1, it was on average 2 times and for cluster 2, on average 4.8 times. On the other hand, votes for Republican in the last 8 years, for cluster 0, it is about 2 times out of 8, for cluster 1, it is 6 times out of 8, and for cluster 2, it is 3.18 times out of 8 on average. This suggest that, cluster 0 tend to be on the average salary, which is the same for the republican salary in cluster 1, democratic party membership, voted on average 6 out of 8 times for democrats and voted about 2 times for republicans in the last 8 years on average. For cluster 1, they tend to have similar salary compared to cluster 0, but mostly on average republican party membership, and vote about 2/8 times democrats and 6/8 times on average republicans. Lastly, for cluster 2, they tend to have very high salary on average, and they tend to lean towards democratic, and voting about 4.8 out of 8 times on average for democrats and voting on average 3.18 out of 8 times for republicans. Their party membership is leaning more towards democrat as well, which might suggest that between the higher earners in urban, there are more voters who lean towards democratic values.

For the rural voters, 90%-95% of the variance explained was on 10 components as well. The AIC scores show similar characteristics as urban voters. The AIC suggested 6 clusters but after fitting the GMM and plotting the centroids, there was again significant overlaps of the cluster's centroid. Using similar reasoning with the urban voters dataset, 3 clusters were used and they were pretty well spread out and the AIC between 3 and 6 weren't too much of a drop.

For the rural voters, age, gender, marital status, rate of voter turnout, most important issue, and votes for other candidates in the last 8 years had very minimal difference in the means between the clusters. Number of children, occupation, standard of living, engagement of campaign had some differences in the mean between the clusters. The biggest differences in the mean between the clusters were salary, party membership, votes democrats in the last 8 years and vote republicans in the last 8 years.

Focusing on the most significant differences in mean between clusters, we can see similar clusters as urban. Cluster 0 and cluster 1 had similar salaries but cluster 2 had a larger mean for salary. Party membership, cluster 0 the mean of the cluster is republican and cluster 1 is democrats. Cluster 2 lean more towards republican since it is below 0.5 but not as much as the urban cluster with large salary. If we look at votes democrats in the last 8 years, we see the mean for cluster 0 to be 2 out of 8 times on average and cluster 1 is 6 out of 8 on average and for cluster 2, about 3.66 out of 8 on

Cluster	Age	Gender	Marital Status	Number of Children	Occupation
0	54.078143	0.573636	1.507458	2.468495	4.368491
1	54.026521	0.569804	1.496035	2.466210	4.661689
2	53.814719	0.577489	1.519481	2.520346	4.559307

Cluster	Salary	Standard of Living	Party Membership
0	50029.160131	0.997685	0.000000
1	49899.602572	0.997844	1.000000
2	500266.003307	1.051082	0.415584

Cluster	Rate of Voter Turnout	Most Important Issue
0	1.090918	2.002734
1	1.076499	1.997528
2	1.083983	2.024242

Cluster	Engagement with Campaign	Votes for Democrat (Last 8 years)
0	1.194381	1.995660
1	1.100172	6.004133
2	1.153247	3.661472

Cluster	Votes for Republican (Last 8 years)
0	5.998626
1	2.002887
2	4.344589

Cluster	Votes for Other Candidate (Last 8 years)
0	1.004066
1	1.001009
2	0.993939

Fig. 6. Mean of Rural Clusters Results

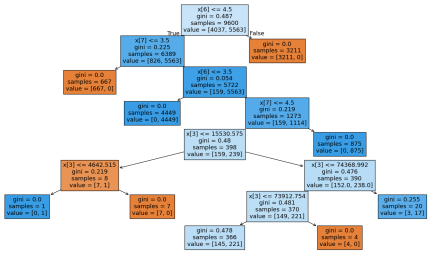

Fig. 7. Decision Tree for Rural Voters

average. Looking at votes for republicans in the last 8 years, we can see cluster 0 voted about 6 out of 8 times on average for republicans and 2 out of 8 on average for cluster 1. For cluster 2, on average it was about 4.34 out of 8. It seems in general, voters mainly vote in terms of party lines, however, we noticed that cluster 2, the high earner voters are split between voting for republicans and democrats. We can see it is almost about 4/8 times on average for both, which could indicate that the higher earners likes some characteristics from both democratic and republican candidates in rural settings. These could be identified as swing voters. The voter membership is close to 0.5 as well but more leaning towards republican in general.

4) Decision-Tree Implementation: We implemented the decision tree model using scikit-learn in Python. Our process involves:

- Complete the data preprocessing.
- Split the data into training (80%) and testing (20%) sets.
- Fitting the model on the training data. This includes removing all qualitative $but not quantitative or boolean$ data from the model.
- Evaluate the model's performance on the test data by developing predictions of the test data using the decision tree, and comparing with the actual test data results.

The model for the rural data was developed with approximately (98%) accuracy. The results of the rural model indicate that approximately (7%) of rural residents are registered or likely Democrats who voted primarily Democrat 4 or fewer

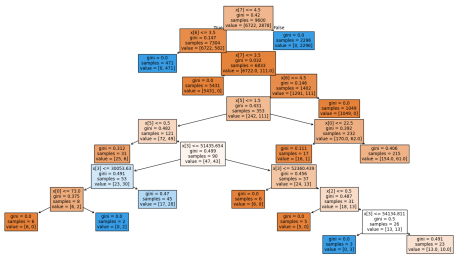

Fig. 8. Decision Tree for Urban Voters

times in the past 8 years. This indicates that Democrats could potentially increase voter turnout by (7%) if they could efficiently encourage registered or likely Democrats to turn out and vote on election day.

The model for the urban data was developed with approximately (98%) accuracy. The results of the urban model indicate that approximately (6%) of uban residents are registered or likely Republicans who voted primarily Republican 4 or fewer times in the past 8 years. This indicates that Republicans could potentially increase voter turnout by (5%) if they could efficiently encourage registered Republicans to turn out and vote on election day.

IV. Comparison

When comparing the two classification models, decision tree and logistic regression, each has its own strengths that make it valuable depending on the analysis goals. The decision tree model stands out with its high accuracy, around 98% for both rural and urban datasets. It's particularly effective in breaking down complex, non-linear relationships and provides clear, actionable insights through its branches. This makes it ideal for identifying specific voter groups that parties could target to boost turnout.

On the other hand, logistic regression, while not as accurate (with accuracy ranging from 59% to 77% depending on the task), offers a unique advantage in its interpretability. By examining the coefficients, one can gain a deeper understanding of how different factors like occupation, salary, and campaign engagement influence voter behavior. This makes logistic regression a strong choice when the focus is on understanding the significance of specific variables rather than just making predictions. Both models have their place, with decision trees being more suited for high-stakes predictions and logistic regression for deeper, more detailed analysis.

V. Future Directions

Some potential directions for the clustering part is that there wasn't a really good way of determining the number of clusters. The AIC score suggested to use 6, however, after plotting the centroids, they show a lot of overlap so it was reduced to 4 but it still had overlap so further reduction was done to reduce it to 3 and this should be fine because the AIC from 3 to 4 was very close. What could be done is to use the silhouette score and to try different number of clusters. This was an issue when we tried because it was very slow, which suggest that it's very computationally costly and was not reasonable to do. We could also try the elbow method,

but both were very slow due to the amount of data. In the future, we could try taking a sample of the data and trying these methods on the sample instead of the whole dataset. For visualization of the GMM, a potential future work could be using t-distributed stochastic neighbor embedding to reduce some of the dimensions for visualization purposes, which was not done due to it being very computationally costly.

For decision tree, to get more insights, we could try an ensemble method like random forest with tuned parameters from gridsearch. Random forest with gridsearch was very computationally costly due to tons of data for both datasets, but it would be a good future direction.

VI. Conclusion

Based on the urban and rural voters, we do see that salary, party membership, votes for democrat and votes for republicans in the last 8 years are the biggest differences between the clusters. The others had little to no differences in mean between clusters. The difference in findings between urban and rural was that in rural, there was a noticeable difference in higher earner voting preferences. In rural, we can see that their voting choices were almost split even on average, and their party membership votes were very close to being split evenly on average but was more leaning republican, whereas in urban the high earners seem to vote more towards democrat. It was an almost 5/8 on average for democrat vs an almost 3/8 for republican. The party membership was also leaning towards democrats more than the rural one.

The logistic regression models provided valuable insights into voter behavior across rural and urban populations, particularly in terms of voter turnout and party affiliation. The models performed reasonably well, especially in predicting voter turnout, with higher accuracy and AUC scores compared to party affiliation predictions. However, the results also highlighted some limitations. The model's performance in predicting party affiliation was less robust, especially in rural areas, where the AUC score was only 0.59. This suggests that while logistic regression can capture some of the underlying patterns in voter behavior, there may be more complex relationships that require more sophisticated models or additional features to fully understand.

The decision tree gave us pretty good insights and results on how we could get 5% or more voters turnout in both urban and rural settings.

VII. Bibliography

- "Summary of Registered Voters and Ballots Cast" *New Jersey Voter Information Portal.* [Online]. Available: https://www.nj.gov/state/elections/election-information-ballots-cast.shtml [Accessed: 24-Jul-2024].
- "USA: Urban Areas" *City Population.* [Online]. Available: https://www.citypopulation.de/en/usa/ua/ [Accessed: 24-Jul-2024].
- "Designated Rural Areas in New Jersey" *Official Site of the State of New Jersey.* [Online]. Unavailable: https://www.citypopulation.de/en/usa/ua/ [Accessed: 24-Jul-2024].