Group Members: Ben Lee (Team Lead), Ander Duong, Namita Hegde, Eric Daniel Tuason, Rui Zhou

Introduction

Voting is a democratic process that gathers individuals to make a collective decision or express their political preferences, which influences the composition of the government. Voting serves as the primary means of ensuring a “government of, by, and for the people.” Yet, voter turnout in the United States is low, hovering around 60 percent in presidential elections and 40 percent in midterm elections. Compared to other countries such as Australia, Belgium, and Chile, voter turnout hovers around 90 percent and 70 percent in developing countries. Many factors contribute to voter turnout, including political interest, electoral competition, voting laws and accessibility, and demographics. For this case, however, only the demographic factor will be studied. There are various demographic factors such as race, gender, education, and income that influence voter turnout and behavior/habits. Setting voting behavior as the dependent variable and race, gender, education, and income as independent variables, models can be built to predict the most accurate fit between the dependent and independent variables.

- Race:
  White people have voted at higher rates than other racial groups. Other racial groups including African Americans, Hispanics, and others have relatively low turnout rates compared to whites.
- Gender:
  Over the years, the turnout rate for women had increased dramatically as a result of women’s suffrage. Since the 1980s, women had voted at about the same rate as men, and sometimes at even higher rates than men.
- Education:
  Citizens with higher education levels are more likely to vote as a result of higher political awareness and knowledge. Education could also motivate individuals to vote by instilling civic values. Exposure to political discussions, rights, and responsibilities can influence voting behavior.
- Income:
  Wealthier individuals tend to vote at higher rates than those with lower incomes. They are likely to have more time and flexibility which allows them to be more involved in politics. Higher-income levels are also often associated with higher education levels, providing them with more opportunities.

Literature Review

Researchers Barber and Holbein analyzed a nationwide dataset of 400 million voting records from the 2014 and 2016 election cycles to determine patterns in voter turnout based on race, age, and political party affiliation. They obtained this data from The Data Trust LLC, which aggregates voter information from all 50 US states and the District of Columbia. The dataset includes variables such as vote history, age, gender, race, geographic location, political party, and other modeled variables. By analyzing this dataset, the researchers aimed to address gaps in understanding voter turnout disparities. They focused on three main aspects: estimating turnout gaps by race, age, and political party; identifying geographic areas with high and low voter turnout (referred to as "turnout deserts"); and exploring the likelihood of individuals with specific social characteristics living in turnout deserts. Comparative to earlier studies, this study methodology enables a thorough examination of the turnout disparities that exist in American democracy.

The methodology used in this paper inspired our research project by highlighting the significance of using comprehensive nationwide administrative data, such as voter files, to analyze voter turnout disparities. It led us to investigate the connections between several variables, such as race, education, and income category, and their effects on political engagement by gaining access to and analyzing a sizable dataset.

The paper from Logan and Darrah used the Current Population Survey (CPS) data from 1996, 1998, 2000, 2002, and 2004, which included a voting and registration supplement. The CPS provided a nationally representative sample that included all major racial and ethnic groups. This dataset allowed for an examination of voting and voter turnout based on race, age, education, and residential stability. They wanted to show racial/ethnic disparities in social-economic resources and rootedness (age, education, residential stability). However, the authors did not explain group differences in elections. This research methodology led us to incorporate contextual variables based on income level into our model and provided insights into the impact of living quality factors on political participation.

The article "Why Do People Vote? A Psychological Analysis of the Causes of Voter Turnout. Journal of Social Issues" written by Harder and Krosnick gave us a good insight into which voter description category or categories could best predict voting frequency and create accurate models to map independent variables (education, race, gender, income level) to the dependent variable (voter frequency). The article explored the impact of independent variables, including education, race, gender, and income, on voting behavior. It specifically highlighted the influence of education and income on voter turnout. According to the article's conclusion, individuals with higher educational levels compared to their community or age group were more likely to vote. Additionally, voter turnouts tended to be greater among the wealthy.

Dataset Description:

The Voting Habits dataset is a comprehensive collection of data that provides valuable insights into the voting behaviors of individuals across different countries. This dataset, last updated in March 2023, is a reliable and highly regarded source of information, boasting a quality rating of 10.00, which was one of the primary reasons we chose to work with it. The dataset contains four independent variables, which are education, race, gender, and income, and one dependent variable: voting habits.

Regarding education, the dataset classifies individuals into three categories based on their highest education level: college, high school or less, and some college. This variable enables researchers to examine the influence of educational background on voting preferences and behavior. The race variable captures the racial composition of voters, including White, Black, Hispanic, and other/mixed communities. This information facilitates an exploration of how race may intersect with political attitudes and electoral choices. The gender variable differentiates between male and female voters, allowing researchers to analyze gender-based differences in voting habits and preferences. The income variable provides insights into the income distribution of voters, categorizing them into four brackets: less than $40K, $40-75K, $75-125K, and $125K or more. This variable is particularly useful in understanding the relationship between socioeconomic status and voting patterns.

Voting habits are the only dependent variable in the dataset, which characterizes the frequency with which individuals tend to vote: sporadic, always, or rarely/never. This variable offers valuable information about voter engagement and can be used to explore factors that may influence voter turnout.

The dataset was compiled from surveys and polls conducted in various countries, ensuring a diverse and representative sample. It received a 100% rating in completeness, credibility, and compatibility, indicating that it is a reliable and comprehensive resource for analysis.

Proposed Methodology:

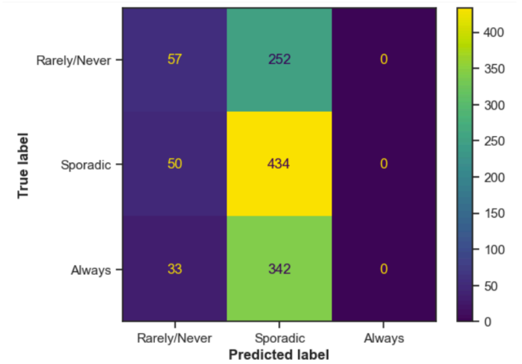
Identification of voter type is a multi-classification problem. Since all of the data is categorical, we decided to create multiple models that are good at processing categorical data. An artificial neural network (ANN), support vector machine (SVM), Naive Bayes Classifier (NBC), and decision tree were all created to determine the best model type for our problem. We ended with 4 different models with only one chosen and fully implemented into the web application. Each model was created using the Sci-kit learn library commonly used in machine learning.

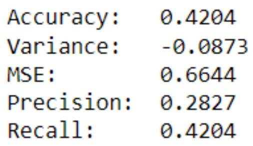
Once our decision tree classifier model was fine-tuned by grid search, we created a web-interface to display our project. The web-interface was created using the Flask library. The website displays the results of the decision tree model, including the confusion matrix, accuracy, and other metrics for measuring the performance of a model. Also included is an interactive portion in which a person can input their demographic data and our model will then predict the frequency of their voting.

The dataset required preprocessing for use within our models and we used the pandas library for this purpose. Though all of the data is categorical, it is not all of the same type. The education, income, and voter category data fields are all ordinal categorical variables; whereas, the race and gender data fields are both nominal categorical data. For the ordinal data, each category has been mapped to a number from 0 to n-1, where n is the number of categories. Namely: for education, the higher the level of education the greater the number; for income, the higher the income the greater the number; and for voter category, the higher the frequency of voting the greater the number. For the nominal categorical data, one-hot encoding was used since these variables have no possible ranking.

Experimental Results:

Support Vector Machine:



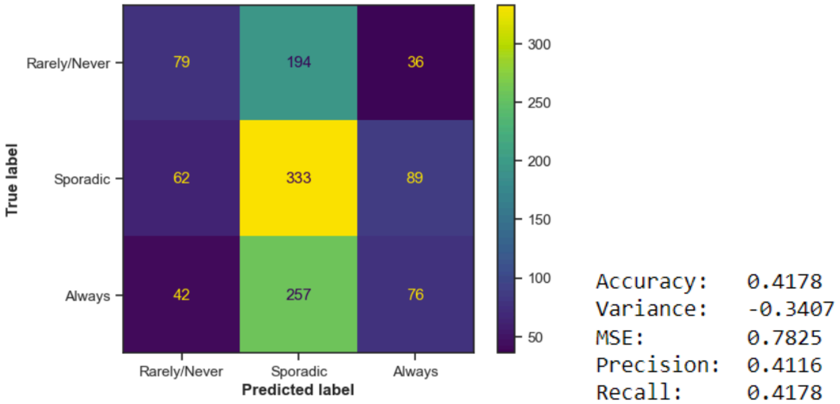


```
Accuracy:   0.4204
Variance:   -0.0873
MSE:        0.6644
Precision:  0.2827
Recall:     0.4204
```

Figure 1. Testing results of SVM

For SVM, the Sklearn library's SVC function was used to create the SVM models. Each possible kernel was tested with its default values. The best kernel for SVM was RBF, which had an accuracy of 42 percent. The next best was linear with an accuracy of 41 percent while all other kernels were below 40 percent. The RBF kernel was tested further by changing the parameters of the model. However, any alteration to the default parameters resulted in lower accuracy. The major issue with this model is that it never predicts that a person always votes, and incredibly rarely will it predict that a person rarely/never votes. Attempts to balance the dataset resulted in lower total accuracy of 38 percent.

Naive Bayes:



True label \ Predicted label	Rarely/Never	Sporadic	Always
Rarely/Never	79	194	36
Sporadic	62	333	89
Always	42	257	76

```
Accuracy:   0.4178
Variance:   -0.3407
MSE:        0.7825
Precision:  0.4116
Recall:     0.4178
```

Figure 2. Testing results of Categorical Naive Bayes

For the Naive Bayes classifier, the Sklearn library's CategoricalNB function was used to create the Naive Bayes Model. Only the alpha parameter altered the model's performance. However any alteration to the alpha parameter either had no effect with a lowered alpha, or a lowered accuracy with a higher alpha. Therefore the best performance was found using the default parameters as defined in the Sklearn library. Similarly to the other models presented in this paper, the model mainly predicted that a person votes sporadically over any other classification.

Neural Network:

The third model we attempted to use was a neural network. Specifically, we attempted to implement an MLP (Multi-Layer Perceptron) Classifier. As we learned in class, this type of feed-forward neural network is organized in the form of several hidden layers, each containing several nodes, which connect the input to the output. The neural network works by passing inputs from each of the input layer nodes through each of the nodes in each hidden layer until it reaches the output layer. At each node, the input is multiplied by the weight of the edge between the nodes and passed to an activation function. The model weights are adjusted during the training process through backpropagation and gradient descent which improves the accuracy.

When implementing an MLP Classifier for our dataset, we used the Scikit-learn MLPClassifier feature. We attempted several different approaches. First, we attempted to use an MLP Classifier with two hidden layers, the first with 12 nodes and the second with 3 nodes and a logistic activation function in each node. Our model used stochastic gradient descent and had a learning rate of 0.3 with 500 epochs. This model achieved 41.78 percent accuracy.

We wanted to attempt to maximize the accuracy, so we tried a few different MLP Classifiers to see if they could achieve higher accuracy. First, we used the same MLP Classifier but used MinMaxScalar on our data to achieve feature scaling. This brought our accuracy down to 39.90 percent, so we did not use this.

Then, we attempted to use the same MLP Classifier with a different train-test ratio to increase the accuracy. We tried a 90:10 train-test split, which produced an accuracy of 43.84 percent. This was an improvement over our original model, but we wanted to try one more train-test ratio. We tried using the same MLP Classifier on a 70:30 train-test split and this produced an accuracy of 43.92 percent.

Since this was the best accuracy we had found so far, we decided to adjust other aspects of the MLP Classifier model while maintaining a 70:30 train-test split. First, we tried different numbers of nodes in each hidden layer. We made one MLP Classifier with 8 nodes in the first hidden layer and 4 nodes in the second hidden layer, which achieved an accuracy of 43.8 percent. We then made another MLP Classifier with 16 nodes in the first hidden layer and 8 nodes in the second hidden layer, which achieved an accuracy of 41.92 percent. Finally, we attempted an MLP Classifier with 32 nodes in the first hidden layer and 16 nodes in the second hidden layer. This model achieved an accuracy of 42.66 percent. Evidently, our first model with 12 nodes in the first hidden layer and 3 nodes in the second hidden layer still had the highest accuracy.

We then attempted to adjust the activation function. When our model used a linear (identity) function, it achieved a 41.8 percent accuracy, while a hyperbolic tangent (tanh) function achieved a 43.18 percent accuracy. We also attempted to use a Rectified Linear Unit (ReLU) function, which gave us a 42.95 percent accuracy. Since logistic still achieved the best accuracy, we kept this as our activation function.

The final adjustment we attempted was to try different learning rates. When the learning rate was 0.1, it achieved an accuracy of 44.37 percent. When the learning rate was 0.2, it achieved an accuracy of 43.23 percent. When the learning rate was 0.4, it achieved an accuracy of 43 percent. The best accuracy was achieved when the learning rate was 0.1.

Thus we found our optimal MLP Classifier Neural Network Model to be used with a 70:30 train-test split dataset, having two hidden layers, the first with twelve nodes and the second with three nodes. The model used a logistic function as its activation function, had a learning rate of 0.1, used stochastic gradient descent, and had 500 epochs. This model achieved 44.37 percent accuracy.

Random Forest:

The fourth model we attempted to implement was a random forest model. Random forest models work by taking random subsets of the data, building a decision tree for each subset, and using the combination of decision trees to make a final prediction of the target value.

We implemented our random forest model using the Scikit-learn RandomForestClassifier feature. We attempted several versions of this model to achieve the highest possible accuracy. The first iteration was simply using our 80:20 train-test split. This model achieved a 41.01 percent accuracy. We then attempted to use a 70:30 train-test split dataset, which achieved a 41.4 percent accuracy. Finally, we attempted to use a 90:10 train-test split dataset, which achieved a 42.29 percent accuracy. The 90:10 train-test split produced the best results in our random forest model.

Thus we found our optimal Random Forest Model to be used with a 90:10 train-test split dataset. This model achieved 42.29 percent accuracy.

Final Model - Decision Trees:

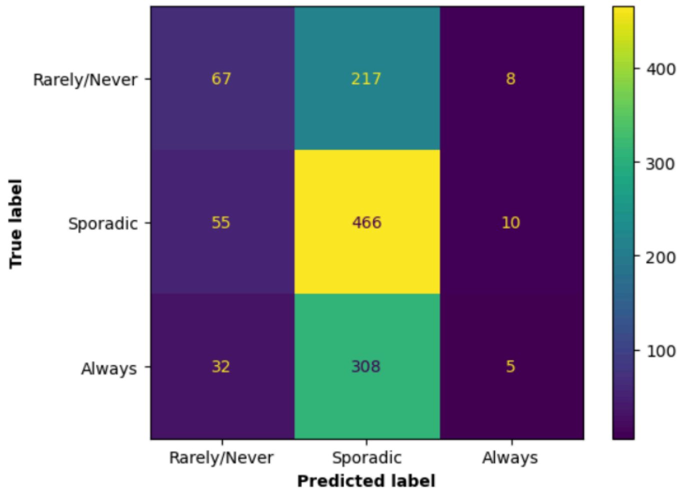


Figure 3. Decision Tree final model results

Overall, our highest accuracy was achieved by using a Decision Tree Classifier Model that relied upon Grid Search hyperparameter tuning. To implement this model, we made use of the Scikit-learn GridSearchCV and DecisionTreeClassifier features.

First, we set up an array of parameters to pass through the Grid Search CV hyperparameter tuning in an attempt to maximize accuracy. The parameters we included were the maximum tree depth, the minimum number of data instances needed at a node to consider further splitting and the minimum number of data instances required at each leaf node.

We then created an instance of the decision tree classifier. We made sure to adjust class weights to place emphasis on the non-target classes in the classification process. We then passed the array of hyperparameters and the decision tree to the Grid Search CV hyperparameter tuner. We set the cross-validation value to 10, which implemented a 10-fold cross-validation.

Once we implemented this, we were able to achieve an optimal decision tree classifier model. Our model has an accuracy of 46.06 percent, with a variance of -0.1362, a mean squared error (MSE) of 0.6421, a precision of 0.3868, and a recall of 0.4606. In comparison to all our other models, this decision tree classifier model clearly had the best performance.

Accuracy refers to the ability of our model to predict the correct label for a given data instance. In the context of our dataset, that refers to the capacity for our model to correctly predict the voting frequency of an individual given their demographic information. The variance of the model refers to the average difference between the expected and predicted values for each of the model's predictions. The mean squared error refers to the average squared difference between predicted and actual values. The precision refers to the model’s ability to correctly identify true positive instances in the data and is calculated by the number of true positives divided by the number of true positives plus the number of false positives. The recall refers to the model's ability to correctly identify true positive instances and true negative instances in the data and is calculated by the number of true positives divided by the number of true positives plus the number of false negatives.

Conclusion/Discussion

Using the dataset nonvoters_dataset our goal was to find which combination of the four independent variables of education, race, gender, and income could be used to best predict voting habits.

Due to the imbalanced nature of the dataset we planned to use methods such as oversampling in order to obtain a more accurate model. However, upon experimentation, balancing methods were found to lower every model's accuracy by approximately five percent. Thus, we omitted mentions of balancing within the experimental results for brevity.

After testing five different models of artificial neural network, support vector machine, naive bayes classifier, random forest, and decision tree, we found that a decision tree with grid search had the best performance. This model produced the most accurate results with an accuracy of 46.06 percent and had variance of -0.1362, mean squared error of 0.6421, precision of 0.3868, and recall of 0.4606.

We also generated a confusion matrix for our model, which allowed us to showcase the predicted and true labels for our test data. Upon analyzing our confusion matrix, we found that the model was skewed toward predicting "Sporadic" as the voter category far more than "Rarely/Never" or "Always". However, when we reviewed the data, we did find that the data itself was fairly skewed toward "Sporadic" voters, so this skewing in our model was not a major cause for concern. Additionally, balancing the training and test data seemed to hurt model performance more than it helped.

Our model results suggest that it's difficult to predict voting frequency just from these four demographic pieces of information due to it's relatively low accuracy of 46 percent. It's likely that more information such as location of these voters may have improved the accuracy of our model.

Works Cited

Barber, M., & Holbein, J. B. (2022). 400 million voting records show profound racial and geographic disparities in voter turnout in the United States. PLOS ONE, 17(6), e0268134. https://doi.org/10.1371/journal.pone.0268134

Logan, J. R., Darrah, J., & Oh, S. (2011). The Impact of Race and Ethnicity, Immigration, and Political Context on Participation in American Electoral Politics. Social forces; a scientific medium of social study and interpretation, 90(3). https://doi.org/10.1093/sf/sor024

Harder, J., & Krosnick, J. A. (2008). Why Do People Vote? A Psychological Analysis of the Causes of Voter Turnout. Journal of Social Issues, 64(3), 525-549. https://doi.org/10.1111/j.1540-4560.2008.00576.x

References

Datasets:
https://www.kaggle.com/datasets/ulrikthygepedersen/voters-and-non-voters