Anna Evans

4/18/2022

CSC-480

Final Research Project

Political Party Prediction with the General Social Survey and Machine Learning Classification

Introduction

For decades political scientists, campaign staffers and their candidates, and others working in political spaces in America have worked to understand what factors lead an individual to affiliate with a certain political party through polls, surveys, and studies. This information is useful to them for a variety of purposes such as understanding human behavior, predicting which voters in a given area are likely to vote or donate money to a specific political campaign, or predicting the outcome of elections. In the current state of the field, political alignment prediction has been moving forward, but there are still deficiencies. Most polls for major races have a hard time decisively predicting a winner due to large margin of errors. One prime example of this is the large number of polls that did not predict Donald Trump being the winner of the 2016 presidential election. Political campaigns are also forced to spend most of their time compiling contact lists for fundraising and canvassing targeting. Voters are usually grouped into voting blocs based off of demographics such as voters who are young women, voters who are rich men, or middle-class voters. Political campaigns will strategize to target these blocs to get out the vote or political scientists might use these demographics to predict the outcome of races. This study will be useful to determine if a collection of demographic factors

will help determine political affiliation or if there is more about voters that needs to be taken into consideration.

The gap in understanding that exists is something that can be explored with machine learning. In the current field, there are a limited number of published studies that utilize machine learning models to predict voter affiliation in America based off of demographic and survey data. Two studies that will be looked at in-depth are "Predicting Political Party Affiliation" by Mathew Arndt and Curtis Miller of the University of Utah and "The Divided (But Not More Predictable) Electorate: A Machine Learning Analysis of Voting in American Presidential Elections" by Seo-young Silvia Kim of American University and Jan Zilinksy of New York University.

Arndt and Miller's research analyzes voter data for every election year from 1948 to 2014 in order to attempt to predict their political party. Their data comes from the American National Election Studies (ANES) Time Series Cumulative Data File which asks questions about voters' beliefs and tracks their voting pattern in America. Arndt and Miller selected 26 variables that they determined would be most useful in terms of predicting political party affiliation and did not have over 30% missing data leading to 47,611 total observations. The target variable they chose was political party and they removed all observations that were "Independent" or "Other" in order to classify only Democrats and Republicans in a binary approach. They then separated the data into training and testing sets and chose four classifiers to run cross-validation on to determine which classifier had the highest accuracy. The four classifiers used were decision trees, decision forests, naïve Bayes, and logistic regression. Their accuracy rates were 65.19%, 65.37%, 66.98%, and 69.44%, respectively. Seeing that the decision forests had the highest cross-validation accuracy, Ardnt and Miller tuned the hyperparameters and got a resulting

accuracy of 68.75% on the testing set, which was close to the accuracy received on the training set. Overall, the authors concluded the study unsatisfied with the accuracy of their best model. For future studies they would have considered including all variables in the label space instead of throwing out independents and third-party voters. They also noted that they would like to experiment with neural networks as classifiers and also use a survey different from the ANES dataset.

The next study by Kim and Zilinsky similarly applies machine learning models to understanding political behavior of American voters. This study seeks to understand if partisan and demographic sorting are present and predictable in the American electorate. In other words, they want to know if belonging to a particular demographic group can predict voting decisions with a high accuracy. Kim and Zilinsky used public opinion survey data from 1952 to 2020 from the ANES, Cooperative Congressional Election Study (CCES), and the University of California Los Angeles' Nationscape surveys. The features used were only demographic variables including age, gender, race, education, and income. The target variable was presidential vote choice which was self-reported. Only those who voted for either a Republican or Democratic presidential candidate were included in the data. In the paper, the authors state that the use of a collection of surveys to draw the data from will make for a model that has better and more honest performance because of the variety of variables and interactions that can be evaluated and a less chance for overfitting. Kim and Zilinksky chose to use random forests to test the accuracy of predicting voting decisions. The data was split into training and testing sets in an 80:20 ratio and the best model was chosen based off of its performance from the training set after tuning the hyperparameters. Once all of the experiments were conducted and the best model was chosen, the final accuracy of the random forests model for predicting presidential voting choice based off

of demographics alone was about 63.5%. While this is not an exceptionally high accuracy, the authors do note that it is better than random guessing. The researchers believe the results also suggest that while the electorate has become more partisan over time, voters are likely not highly partisan along demographic lines.

While these studies contrast in their approach to answering their respective research questions, they will both serve to inform this research in several useful ways. As stated earlier, the literature in the field of machine learning and political science is limited. While there are several studies that use natural language processing to predict the political ideology of speeches made by politicians or tweets made by twitter users, the literature included above seemed to be some of the few studies that attempt to predict political party affiliation of American voters using demographic survey data. In order to fill the gap in the existing literature, I plan to use demographic variables from the General Social Survey (GSS) data which neither study above included in their research in order to predict American's political party affiliation. I also plan to use a similar approach to both of these studies by using a selection of classifiers, tuning the hyperparameters in order to get the best model, and then obtaining the highest accuracy possible. This will allow us to test if applying existing models to a novel dataset of demographic features will lead to a higher accuracy.

Methods

The data I will be using comes from the General Social Survey (GSS) which is created and administered by the National Opinion Research Center at the University of Chicago (NORC) each year. The GSS is an extensive survey with over 5000 variables on attitudes and beliefs on various social and political issues among Americans 18 years and older. The data available has

been collected each year since 1972. For this study, I will be using the 2021 dataset consisting of 815 variables in order to capture the most up-to-date American voter demographics.

For the feature selection step in the data collection stage, I needed to narrow down the dataset from 815 features to a smaller, more meaningful collection of features that relate to the aims of this project. Since the project is aimed at predicting political party affiliation based solely off of demographic factors and not attitudes, I went through the list of variables and chose features that were not attitude-based questions and related most to the realm of politics and what can contribute to voters' decisions at the ballot box. Twenty-nine variables were initially selected and were then each evaluated for amount missing data. Variables where over 10% of the values were missing or NAs were not included in the feature selection process. The final dataset had ten features which are listed in a table in Figure 1. In total, there were 3,069 observations in the dataset.

Figure 1

Feature ID	GSS Survey Question
Age	Respondent's age
Educ	Respondent's highest year of school completed
Sex	Respondent's Sex
Born	Were you born in this country?
Degree	Respondent's highest degree earned
Incom16	Thinking about the time when you were 16 years old, compared with American families in general then, would you say your family income was: far below average, below average, above average, or far above average.
Attend	How often do you attend religious services?
Marital	Are you currently married, widowed, divorced, separated, or have you never been married?
Relig	What is your religious preference? Is it Protestant, Catholic, Jewish, some other religion, or no religion?

Finalter	During the last few years, has your financial situation been getting better, worse, or has it stayed the same?
Wrkgovt	Are/were you employed by the government? (including federal, state, or local government.)
Partyid	Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?

The data cleaning, preprocessing, and feature engineering stages involved filling in missing data, converting categorical variables, and converting the target variable in order to normalize the data. For missing data, the most frequent value, or mode, of each feature was filled in to replace NAs. All of the data with 'Yes' or 'No' responses were converted to binary 0 and 1 values. Age was also converted into age brackets roughly according to generation in ascending order as follows: 1 (ages 18 to 25), 2 (ages 26 to 41), 3 (ages 42 to 57), 4 (ages 58 to 76), and 5 (ages 77 and older). There were several features that were categorical and were either ordinal or nominal. The ordinal variables which have a natural order were not converted since their values were already in their properly ranked orders. The ordinal variables were degree, incom16, and attend. However, nominal variables which have categories that do not have an order or rank were converted to one-hot encoded variables. This means for each of the possible responses for these features, a new column was created, so whichever choice was a given case's response, the cell for that feature response would have a value of 1 and rest would be 0. These nominal categorical features that were converted were marital, relig, and finalter. The Party ID variable has four possible responses in the original GSS survey which are Democrat, independent, Republican, and other. Since the goal of this study is to create a binary classification model of Democratic and Republican voters, the independent and other targets were filtered out of the data set leaving us with only observations that had Democrat and Republican targets. The value counts of the two

target classes were unbalanced at 1,834 Democrats and 1,235 Republicans. It is important to have a balanced dataset to avoid distribution bias in predictive ability when it comes to training the models (D'Angela, 2021). In order to create balanced distribution of targets, we used the synthetic minority oversampling technique (SMOTE) to oversample the minority class. SMOTE is unique and especially useful because instead of simply duplicating existing rows which could lead to overfitting, it generates additional samples to create balance and avoid model overfitting. After balancing the data, the number of Democrat and Republican targets were balanced at 1,834 and the total number of observations in the data sets was 3,668. Since Party ID is the target feature in this study, this variable was separated from the dataset and put into its own data frame leaving us with one data set containing twenty-nine feature columns and 3,668 rows, and a second target data set containing the target column and 3,668 rows. This completed the data preprocessing, cleaning, and feature engineering stage of the machine learning pipeline.

Four models were selected to test and compare the accuracy on the data. In order to obtain the highest possible accuracy score, a selection of models were selected to compare accuracy rather than starting with only one because that would be a limited approach and potentially introduce a preference bias. The four classifiers chosen were Decision tree, Multi-Layer Perceptron, Naïve Bayes, and Random Forest. These were chosen because they encompass a selection of supervised learning algorithms and neural networks that can be used for classification.

The decision tree provides a simple learning method with a predictive modeling approach. They are a supervised learning method commonly used for classification but can also be used for regression. Decision trees take target labels and feature variables. The decision tree is a series of nodes connected by branches where each node represents a class label and the

branches represent combinations of features that best split the data and help make the best predictions (Mitchell, 1997).

The Multi-layer Perceptron (MLP) is a neural network. They are networks with input and output layers and can be increasingly complex through several layers hidden layers within that have various activation functions. The layers of MLPs learn through searching the space of the matrices and applying weights until the combination of weight values that most closely is able to linearly separate positive and negative examples is reached. MLPs are versatile in that they can be used for binary classification, multi-class classification, and regression problems (Mitchell, 1997).

Naïve Bayes is a supervised learning algorithm and probabilistic classifier that uses conditional probability to make predictions. Naïve Bayes follows the Naïve Bayesian approach which is a tractable notion that comes with an assumption of conditional independence meaning each node is assumed to be independent of its non-descendants given its immediate parents (Mitchell, 1997).

Random Forest represents an ensemble learning method that uses a modified tree learning algorithm that consists of many decision trees. In the learning process, this algorithm selects the best features from a random subset of features to make the its predictions. Using randomness allows for each individual tree to be an uncorrelated collection of trees which creates a strong prediction as a whole (Flach, 2012).

As we can see, each of these models brings their own strengths and represent a variety of methods that give us a good cross-section of models to examine when evaluating the accuracy for each. In order to compare the accuracy of each of the models, I will use 10-fold cross-validation. Cross-validation is a technique that assess how well a model performs on data that is

new to the model and gives an indication of how well the model can generalize to new data (Gupta, 2017). It's useful here because it allows us to obtain the accuracy of each model using the entire dataset. The dataset used here is relatively small, so avoiding partitioning out any data for this step will be valuable. After cross-validation, the data will be split into training, validation, and testing sets in a 60:20:20 ratio. The model that returns the highest accuracy after cross-validation will be further trained on the training data, tested on the validation data until the hyperparameters are optimally tuned to yield the highest possible accuracy, and then tested on the test data to get a final accuracy score. The final model will be evaluated using the following metrics: ROC curve, confusion matrix, accuracy score, F1 score, precision score, and recall score. These metrics will be valuable in the information they provide about the performance of the model, specifically in what they predict well and what they do not. Precision score represents how well the model correctly predicts positives of the class out of all of the positive predictions it makes. The recall score represents how well the model correctly predicts positive targets and negative targets in the dataset. The F1 score weighs both precision and recall and is useful for looking at optimization with either precision or recall. The ROC curve will show the correctness of the model's predictions. A more accurate model will display a curve that is closer to the top left of the graph. Finally, the confusion matrix will show the percentage of true positives, false positives, true negatives, and false negatives in order to again assess the performance of the model as well as assess the ratio of true negatives and true positives (Kanstren, 2021). I will be using the Python programming language and the Sci-kit Learn package containing the various classifiers and metrics necessary to execute the experiments. The code where all of the cleaning and experiments were executed is available in the attached "Final Research Project Code" PDF file.

Results

The accuracy of each classifier was determined through cross-validation. All of the accuracy results are displayed in Figure 2 below. The random forest classifier had the highest accuracy score of the selection of models at 71% with a standard deviation of .028. The Multi-Layer Perceptron classifier was very close behind in accuracy at 70% with a standard deviation of .15. The Decision Tree classifier and Naïve Bayes did not break the 70% accuracy threshold at 67.62% with a standard deviation of .14 and 65.27% with a standard deviation of .06, respectively. It would make sense that the Random Forest classifier worked best compared to the others because it works faster and more efficiently with high dimension data compared to decision trees. A table of the accuracy scores for each model are illustrated in the table in Figure 2 below.

Figure 2

Learning Classifier	Accuracy	Standard Deviation
Decision Tree	67.62%	.14
Multi-Layer Perceptron	70%	.15
Naïve Bayes	65.27%	.06
Random Forest	71%	.028

Once the Random Forest classifier was found to have the highest accuracy of the rest, the base Random Forest model was trained on the training data and tested on the validation data. The result of this was a 71.39% accuracy score. In order to tune the hyperparameters, we used a Randomized Search CV to find the best features in the most efficient way possible. The Randomized Search CV executes a random search over all of the given parameters where each

combination of settings is sampled from a distribution. The search is cross-validated and the combination of parameter settings that has the highest accuracy score is returned. This method was used as opposed to Grid Search CV because a grid search is too costly since it conducts an exhaustive search over all of the parameter values. Therefore, the Randomized Search was much more efficient in terms of run time and cost. The parameter grid given to the Randomized Search CV is seen in Figure 3.

Figure 3

```
{'bootstrap': [True, False],
 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
 'max_features': ['auto', 'sqrt'],
 'min_samples_leaf': [1, 2, 4],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}
```

After fitting a random selection of the parameters to the Random Forest classifier, the Randomized Search CV returned a list of best parameters. The best parameters list is shown in Figure 4.

Figure 4

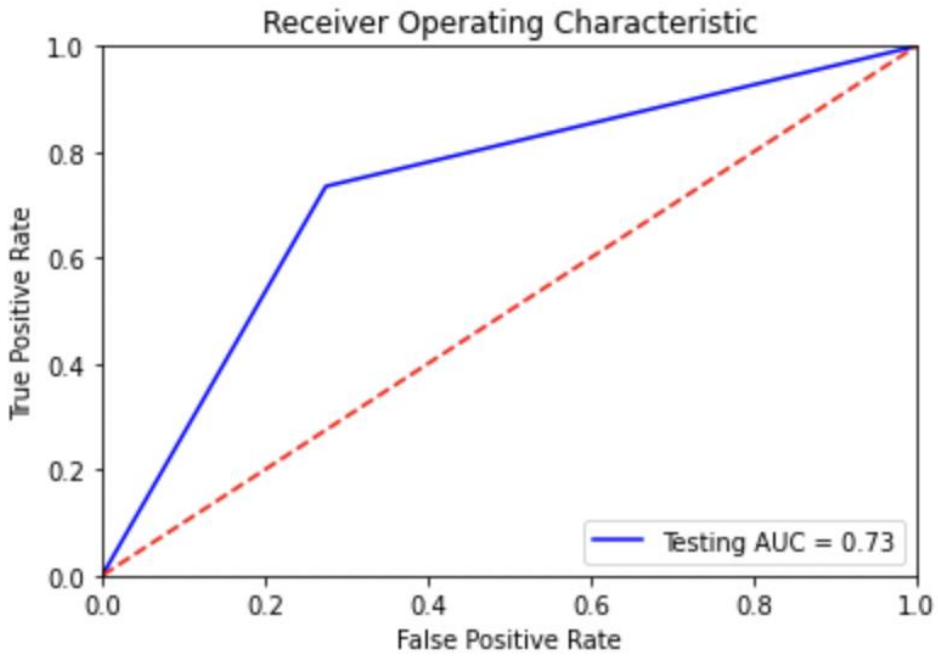
```
{'bootstrap': False,
 'max_depth': 80,
 'max_features': 'sqrt',
 'min_samples_leaf': 1,
 'min_samples_split': 10,
 'n_estimators': 1000}
```

The final model with the best parameters as determined by the Randomized Search CV was created with the parameters displayed in Figure 4 and fit to the training data and tested on the validation data. This model had an accuracy of 72.48% on the validation data. This score is 1.09% better than the base Random Forest model. Seeing that this model performed well after hyperparameter tuning, the model was tested on the testing data in order to obtain the final

accuracy score and metrics. This resulted in an accuracy score of 73.16% which is roughly the same score as the accuracy on the validation data.

A plot of the ROC curve is shown below in Figure 5. The curve reiterates the accuracy of the model by plotting the true positive rate on the Y-Axis against the false positive rate on the X-Axis. It shows that the model is better than random guessing which is displayed as the dotted red line. The model is more on the conservative side because of its lower ratio in its classification of positive examples.

Figure 5

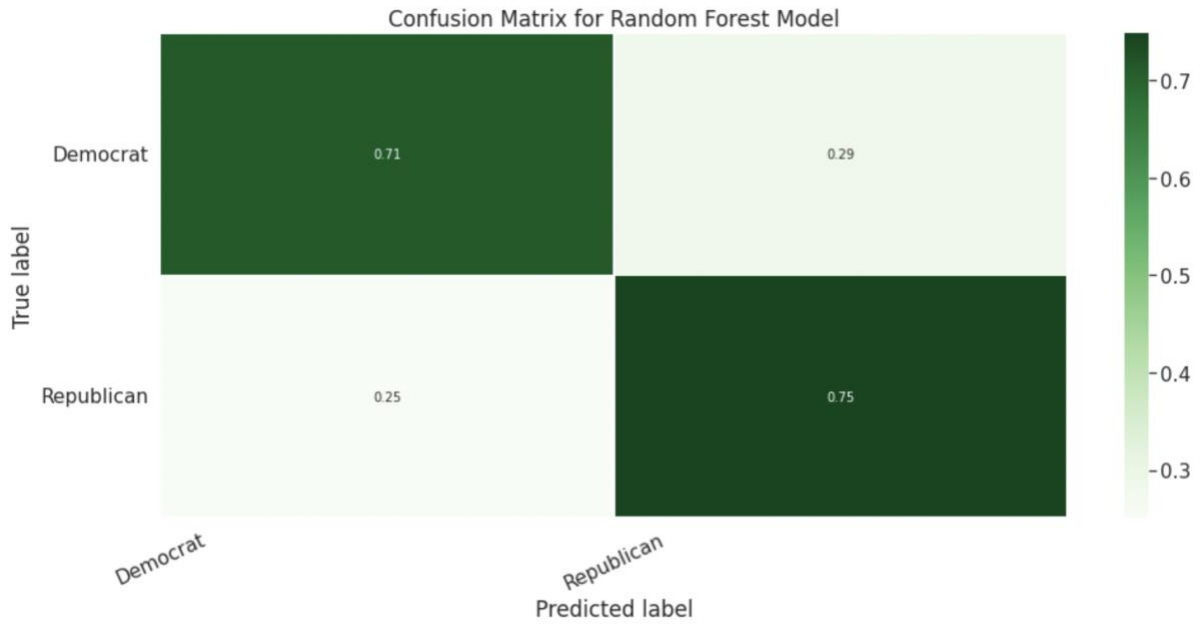


Other metrics assessed were precision score, recall score and F1 score. The model with the best parameters had a precision score of .73, a recall score of .73, and an F1 score of .73. Compared with the metrics from the baseline model which had a precision score of .71, a recall score of .71, and an F1 score of .71, which are all .02 below the tuned model. The similar scores across all of the metrics show that the model classifies an equal amount of false positives as it

does false negatives. Therefore it makes sense that the F1 score, which is a measure that weighs both precision and recall, is roughly the same as precision and recall.

A plot of the confusion matrix is below in Figure 6. Again, it shows that the model predicted roughly equal shares of true negatives (Democrats) and true positives (Republicans). There was a small percentage of 4% more true positives than true negatives.

Figure 6



Another interesting result obtained that is worth mentioning was numerical feature importance. This comes from the decision tree classifier that determined which features were the most important in determining the target class. Four features had an importance of 10% or higher which were age at 10%, incom16 at 13%, educ at 13%, and attend at 15%. This means how often the respondent attends religious services, the highest education level completed by the respondent, the respondent's family income when they were 16 years old, and the age of the respondent were the most important features for determining whether the respondent was a Democrat or Republican as defined by the decision tree classifier.

Conclusion

The results show that our model is roughly 73% accurate in predicting the political affiliation of American voters based off of a selection of ten demographic features from the 2021 GSS data. Although this is not a perfect accuracy score or close to the 90% accuracy range, this model is better than the others that were looked at in the existing literature. Namely the research done by Arndt and Miller which used ANES survey data and a similar machine learning approach and obtained a final accuracy score of 68.75%, about 4.25% less than our model. Perhaps this is because the GSS survey contained different demographic variables that have more influence in influencing political party affiliation. For instance, the GSS had some unique survey questions that were not present in the ANES that the decision tree classifier found to be most important in determining political party affiliation such as family income of the respondent when they were 16 years old.

However, there are clearly still some demographic or attitudinal factors not included in this research that could increase the accuracy of the model in classifying political party affiliation. As was alluded to in Kim and Zilinsky's study, demographic factors alone cannot predict political party affiliation. Nevertheless, the GSS features included here help the model do a better job predicting political party affiliation than existing studies using ANES, CCES, and Nationscape feature data.

Moreover, the results reveal that individuals or groups aiming to either predict voters' choices at the polls, influence groups of people to vote for or donate money to a political candidate, or simply trying to understand American political behavior cannot rely solely on demographics to determine political party choice. For example, some parties may consistently assume they have the vote of a demographic bloc, but this is becoming less of the case as seen in

Donald Trump's unexpected 2016 presidential election victory. As made evident in this article that appeared in the Washington Post, the polls in 2016 suffered many issues in correctly predicting the outcome of the election because the weight of the samples that were usually applied to demographics of voters did not apply anymore (Balz, 2021). Political parties, pollsters, and others need to adapt and take a deeper look at voters beyond their demographics to understand the electorate better in today's society in America.

Future research can explore many different avenues of experiments that were not explored here. For one, it could expand the model to perform multi-class learning by predicting independent and third-party voters in addition to Democrats and Republicans. Although independents and third-party voters make up a small portion of the electorate, they are still important and can be influential in elections, so this information would be useful to various stakeholders. Future research could also experiment with other classifiers such as SVM or look into unsupervised methods such as clustering. There are over 800 other variables that could be used as features for another study, so experimenting with more features or expanding the time frame to look at voter affiliation over time rather than just focusing on 2021 could lead to interesting results.

Overall, I believe that this model has provided some new insight into the field of machine learning and political science. The GSS data proved to provide some unique features that were not present in survey data used in other research models which resulted in a higher model accuracy.

References

Arndt, M., & Miller, C. (2016). Predicting Political Party Affiliation. Retrieved from https://ntguardian.files.wordpress.com/2016/08/cs6350project.pdf.

Balz, D. (2021, July 19). *2020 presidential polls suffered worst performance in decades, report says*. The Washington Post. Retrieved April 25, 2022, from https://www.washingtonpost.com/politics/2020-poll-errors/2021/07/18/8d6a9838-e7df-11eb-ba5d-55d3b5ffcaf1_story.html

D'Angela, A. (2021, February 4). *Why weight? the importance of training on balanced datasets*. Medium. Retrieved April 25, 2022, from https://towardsdatascience.com/why-weight-the-importance-of-training-on-balanced-datasets-f1e54688e7df

Flach. (2012). Machine learning : the art and science of algorithms that make sense of data / Peter Flach. Cambridge University Press.

Gupta, P. (2017, June 5). *Cross-validation in machine learning*. Medium. Retrieved April 25, 2022, from https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Kanstren, T. (2021, May 19). *A look at precision, recall, and F1-score*. Medium. Retrieved April 25, 2022, from https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

Kim, S.-young S., & Zilinsky, J. (2021). The divided (but not more predictable) electorate: A machine learning analysis of voting in American Presidential Elections. https://doi.org/10.33774/apsa-2021-45w3m

Mitchell. (1997). Machine Learning / Tom M. Mitchell. McGraw-Hill.