Donovan Miner

12/12/2023

Data Mining

Dr. Yijun Zhao

Presidential Election Random Forest Model

General Introduction, Features, and Data Collection

As a political science undergraduate student, I was motivated to create a model related to my undergraduate major, and a model attempting to predict presidential elections seemed the most enticing. The model itself does not necessarily predict the outcome of the election, but rather the voting results of individual states during a given election year. The results of the individual states can then be used to determine the hypothetical winner of the election of that year based on the number of district seats assigned to each state, but the focus of the model lies in individual states. The model is a binary Random Forest classifier that labels a given state in a given election year with one of the two current major political parties in the U.S., Democrat or Republican.

The features for the model were drawn from my studies of political science here at Fordham. The initial set of features were to consist of information related to age, income, education, marriage status, religion, patriotism, racial demographics, and population density. As for why, younger generations are typically the more progressive voices in modern politics, causing a correlation in voting pattern based on age. From this, states’ voting patterns in the presidential election would also be correlated with their younger population, or lack thereof. Income, population density, patriotism, education, and religion are all correlated or inversely correlated with the two major parties based on those parties’ respective policy views, and as a result were chosen as features to test as well. Racial demographics were included as many minority populations have been shown to vote at a much higher rate for the modern Democratic party due to its role in the passing of the Civil Rights Act of 1964. Marriage statistics were somewhat of an experimental variable, as it has been shown that married women tend to vote more conservatively than single women. Based on this, it is expected that states with higher marriage rates would also vote for Republican candidates in the presidential elections. Finally, population density was chosen as urban areas align highly with the Democratic party, while rural areas with the Republican party (Pew Research, 2018).

Data was largely collected from the U.S. Census website (United States, 2020). Data was available in the form of the decennial census counts, as well as intercensal estimates . Data was collected for almost all the features of interest, in the form of 5-year age cohorts, median income, educational attainment by population, number of married families, number of veterans, and racial demographic counts. All data was collected at the state level and only included the possible voting population of those at least 18 years of age. However, the U.S. only posts the two most recent census tables online at a time, while the remaining census counts can be found in their online archives in ASCII formats. There is no direct software support for decrypting these files, and the only linked instructions for doing so were for Microsoft Access ’97 and were posted on the U.S. Census’ old website, which no longer exists. As a result, the demographic data before 2000 had to be estimated based on existing and retrievable data. Income estimates were collected from SAIPE (Small Area Income and Poverty Estimates, 2022), and estimates were collected for all but a few years between 1988 and 2022. Religious data is not collected by the federal census, so it was instead collected from the Association of Religion Data Archives and Pew Research (ARDA, 2020) (Lipka and Wormald, 2023). This data was incredibly scarce however, with only 6 data points collected per state. Election results were gathered from the FEC’s archives of federal election results (Federal Communications Commission, 2020).

Completing the Data Set

1988 marked the cut-off for data collection in this project as that is when federal records became scanned copies of paper records, making further data collection much more difficult and time consuming. The first feature to be completed in the data set was median income, as much of this data already existed. The final few data points for each state were filled by interpolating the existing data. This was also used to fill in the few missing data points from the collected Census data.

A graph with orange dots and white text

Description automatically generated

Fig. 1: Interpolated median income (in thousands) for Florida.

The next feature to be completed was total state population. While not a feature in the model itself, total state population was necessary to compute in order to fit features to a Gaussian distribution. The U.S. is disproportionately filled with low and middling population states, with a few extreme outliers, leading to incredibly positively skewed demographic data. When accounting for these populations as a portion of state population, they fit a Gaussian distribution and are easier to compare (see Fig. 2-5). For example, if left in raw counts, or even normalized in raw counts, large youth cohorts in rural areas would not be considered significant by the model, despite possibly making up a large portion of their state population and having more sway over their individual states’ election.

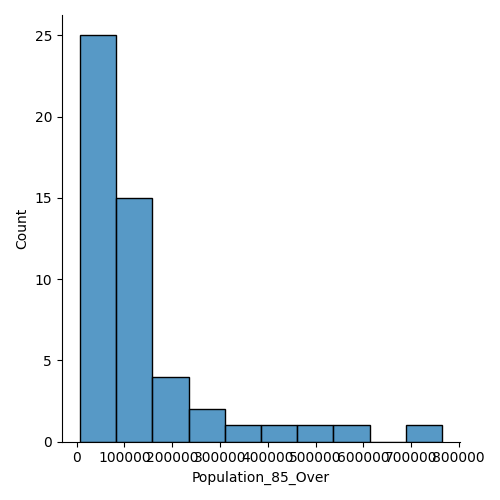
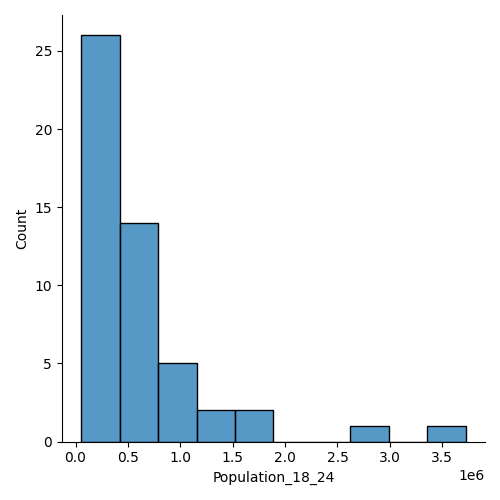


Fig. 2 and 3: Raw population counts for cohorts 18-24 and 85 and over for all states in 2020 (Fig. 2 in millions).

A graph of a number of blue bars

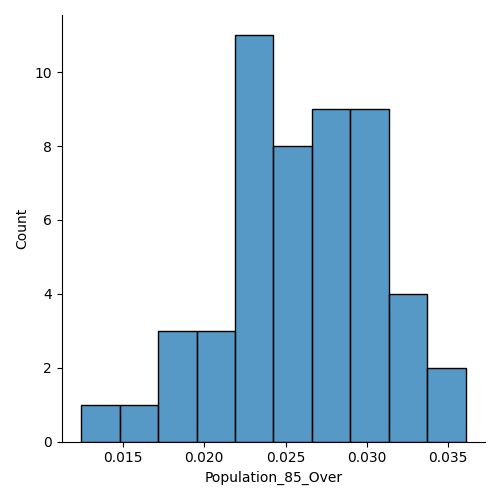
Description automatically generated

Fig. 4 and 5: Cohorts 18-24 and 85 and over as a portion of population for all states in 2020.

As there were no reliable data to begin modeling populations through auto-regression or exponential smoothing, state populations were ultimately modeled using existing patterns of population growth, exponential growth and decay. Based on the average growth rate of an individual state in the existing data, its doubling or halving time was calculated using the rule of 70. The population was then halved, and the start time of the equation taken as the difference of 2022 and the respective doubling or halving time. This technique was chosen over typical regression as it was common for linear or even quadratic regressions to poorly represent population changes in certain states, but even more so in individual age cohorts, as a data point or two out of line with the rest of the series would offset trends and create inaccurate estimates. State population estimates were calculated using the following equations.

EQ1: Pt = P0 \* 2 t / j

EQ2: Pt = P0 \* 0.5 t /j

Where: Pt is the population at *t* years after the starting date.

P0 is the starting population.

*t* is the number of years after the starting date.

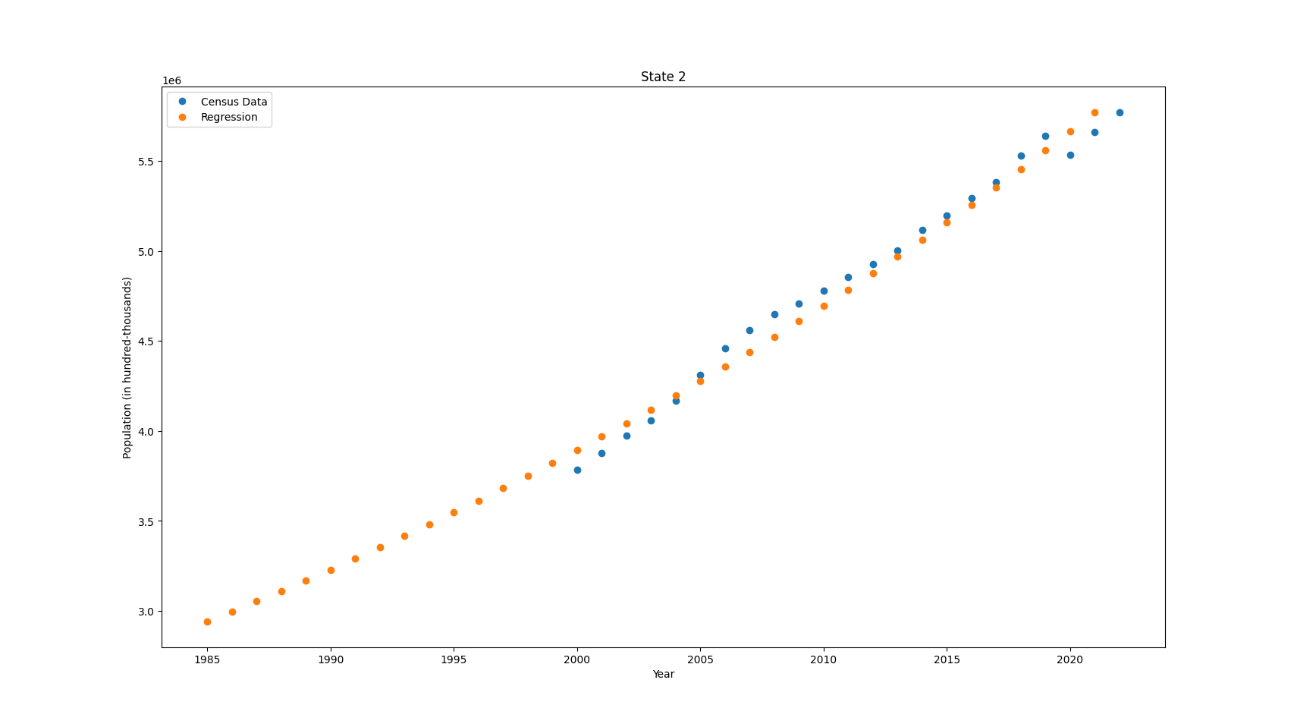
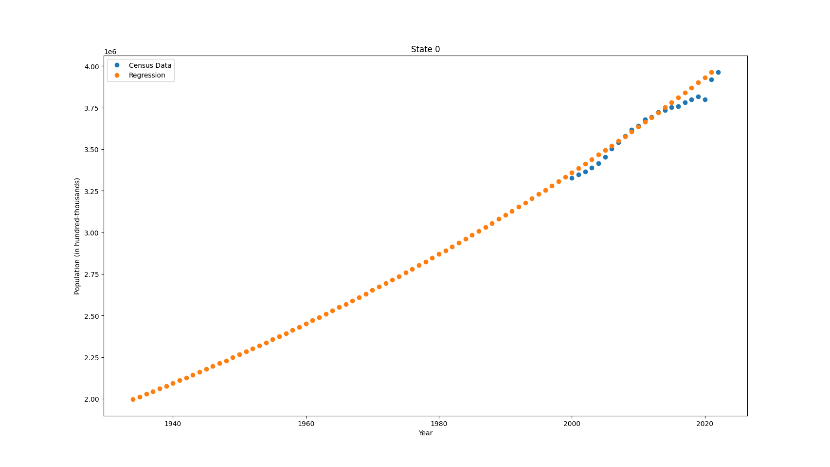
*j­* is the total number of years, or duration of the doubling/halving time. 

Fig. 6: State population estimates for Arizona, based on the above methods.

A graph with orange dots

Description automatically generated

Fig. 7, 8: State population estimates for Alabama, time frames differ between states based on average growth and doubling time. Poorly fitting linear regression for 18-24 age cohort in Alabama.

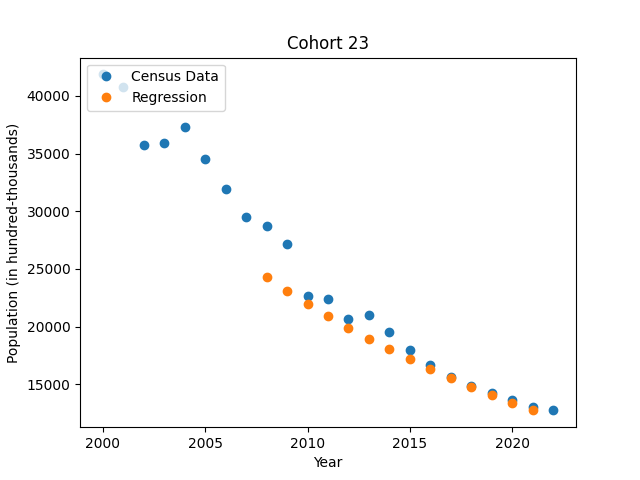
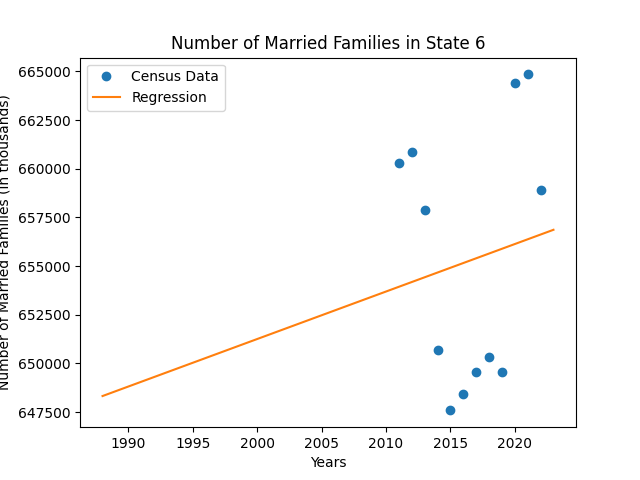
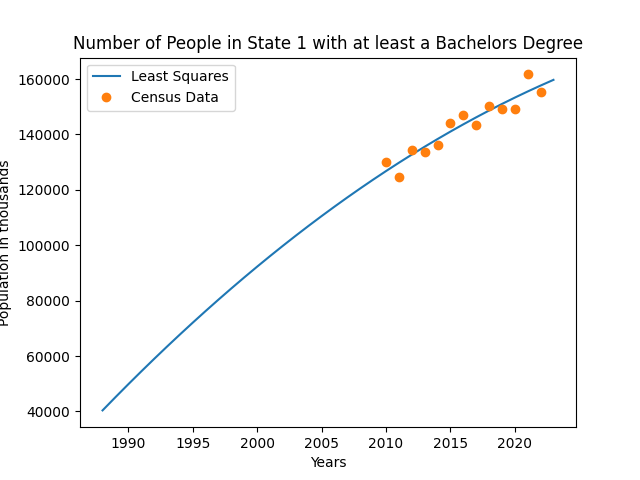
Age cohorts were calculated in the same manner as state populations. However, some age cohorts demonstrated extreme growth or decline, and were unable to produce results before 1988 due to their short doubling or halving times. Others demonstrated oscillating or stagnant growth, typically older age cohorts that remained somewhat constant over time. To differentiate these instances, a population model with a squared Pearson correlation coefficient less than 0.7 (Pauli, et al, 2020), or a mean absolute error (MAE) greater than 12% of the mean population were rejected. The MAE was divided by the mean average population of the cohort to make samples more comparable, while providing insight into individual population models. This measure, however, was still subject to the population range of the cohorts and led to greater errors in smaller populations. This led to the introduction of the additional r2 measure (Pauli, et al, 2020).

Fig. 9: 60-64 age cohort in Alaska that did not produce data to or before 1988.

Cohorts that did not meet these criteria had their missing values filled with random values falling within a Gaussian distribution determined by the mean and standard deviation of the cohort’s existing data (Harris, Millman, van der Walt, et.al, 2020).

The education, marriage, racial demographics, and veteran populations were all modeled linearly, unless the regression gave unreasonable results, such as negative populations by 1988. In these cases, a quadratic regression was fitted to the data (Harris, Millman, van der Walt, et.al, 2020). Once again, a r2 of 0.7 was used as the threshold for a good representation, otherwise values were randomly fitted to the mean and standard deviation.

Fig. 10, 11: Quadratic regression of Alabama’s graduate population, low fit marriage population in Connecticut.

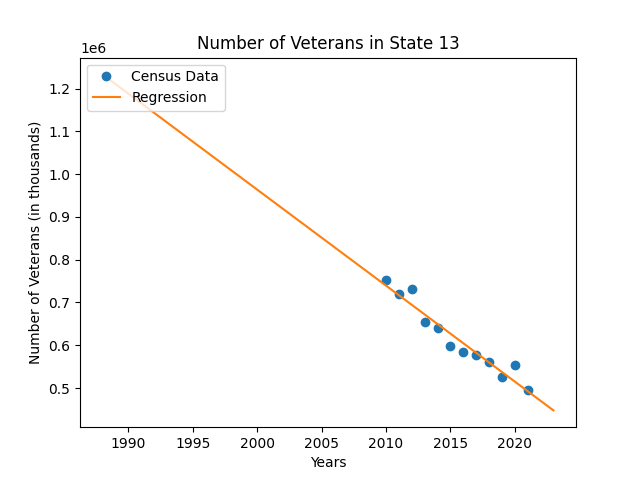
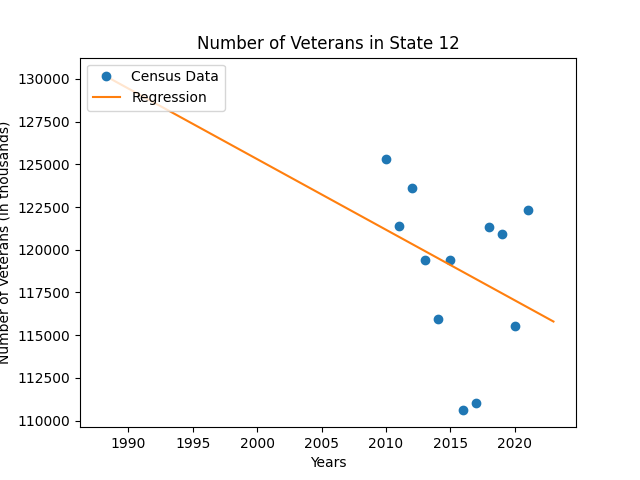


Fig. 12,13: Low fit veteran population in Idaho vs. typical veteran population regression in Illinois

A graph of different colored lines

Description automatically generatedReligiosity proved to be the most difficult to deal with feature. Many states demonstrated oscillating religiosity measures from decade to decade. From my three and half years of political science studies here at Fordham, I had no clear explanation to this pattern which was present in most states. The clearest explanation I had was that religious affiliation had changed over the decades along with and as a result of large-scale social changes. Perhaps as social norms changed

Fig. 14: Oscillating pattern of religiosity measures

or social unrest rose and fell over the decades, so had religious affiliation, and would result in a political indicator. Based on this, the delta between the crest and trough of the religious affiliation data over the decades was measured and tested for correlation with the model labels, but there was none present. Ultimately these series were also filled randomly within a Gaussian distribution. In total 55 series were randomly assigned out of the potential 1, 071 population subdivision. 1 of these series was in the veterans’ feature, 13 in marital status, and the remaining in racial demographics.

Building the Model

Now that the data set has been filled in the model itself can be created. The following features were initially used in the model.

* Age Cohorts, ranging from 18 to 85+, in bins of 5, as portions of state population
* Median Income (Z-Score normalized)
* College graduate population, as a portion of state population
* Married families, as a portion of state population
* Religiosity, as a percent of state population (Min-Max Normalized)
* Veteran population, as a portion of state population
* White, minority, and Hispanic populations as portions of state population
* Population density (Min-Max normalized)

Median income was z-score normalized across election years to also account for changes in inflation. Religiosity was min-max normalized, rather than z-score normalized, to better emphasize differences in religiosity scores, as many clustered around the 30-50% marks. Washington D.C. is a city and not a state, but still a federal district with votes in federal elections, its population density was smoothed from over 8,000 to 1,000 as it made it impossible to smooth other data points though z-score or min-max methods, as the next highest population density was 552 people per mi2. By smoothing this value and using min-max normalization, D.C. still retained the highest population density, but allowed for other states to be normalized.

By simply plugging in the unnormalized values into SK-Learn’s Random Forest classifier (Pedregosa et al. 2011), with an ensemble of 10 trees and a maximum depth of 5, an accuracy of 66% is recorded. Normalizing density, religiosity, and median income resulted in a 5% increase in accuracy. Features were initially tested with a Decision Tree classifier to test their correlation with the label, however, features could only be tested heuristically in this manner, and results were scattered and difficult to assess. In the end, dimensionality reduction and features selection were completed using a Sequential Forward Selection process (Pedregosa et al., 2011). A cross validation of 9 was used in this process, as this was the number of election cycles the data covered, and was reflective of the amount of data the model would be trained and tested on while in use. An ensemble of 100 trees were used throughout this process, and a max depth of 24 (the number of features plus 1) to allow all features and their relationships to be present while reducing dimensionality. The features selected from this process included the following 11 features.

* Age Cohorts (18-24, 35-39, 45-49, 55-59, 75-79, 85+)
* Median Income
* White, Minority, and Hispanic demographics
* Population Density

After reducing dimensionality, model accuracy increased to 82.5%. From this, tree depth was tested, which showed that accuracy plateaued around a depth of 8. The test took the average of ten, 9-fold cross validated accuracy scores (Pedregosa et al, 2011).

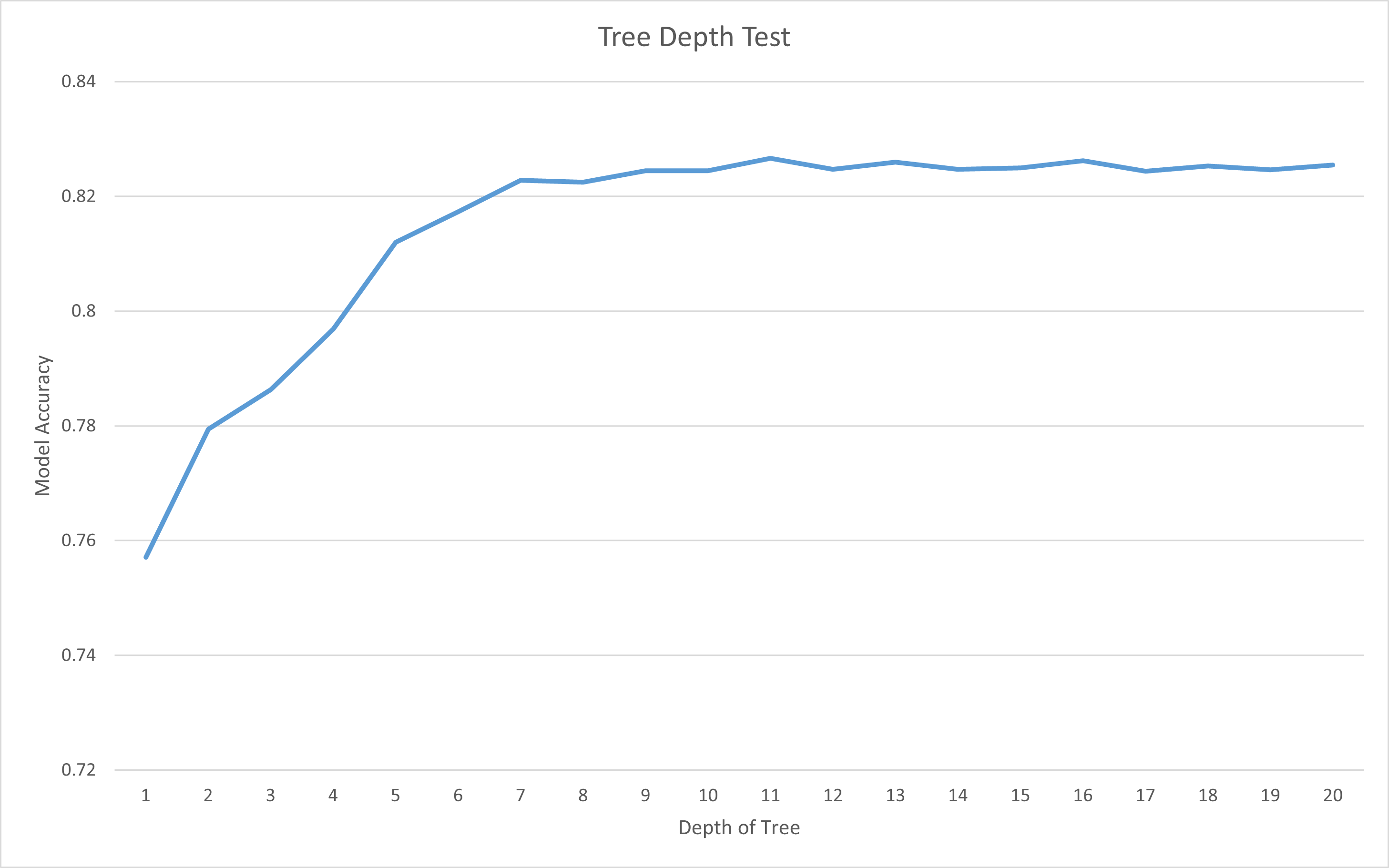
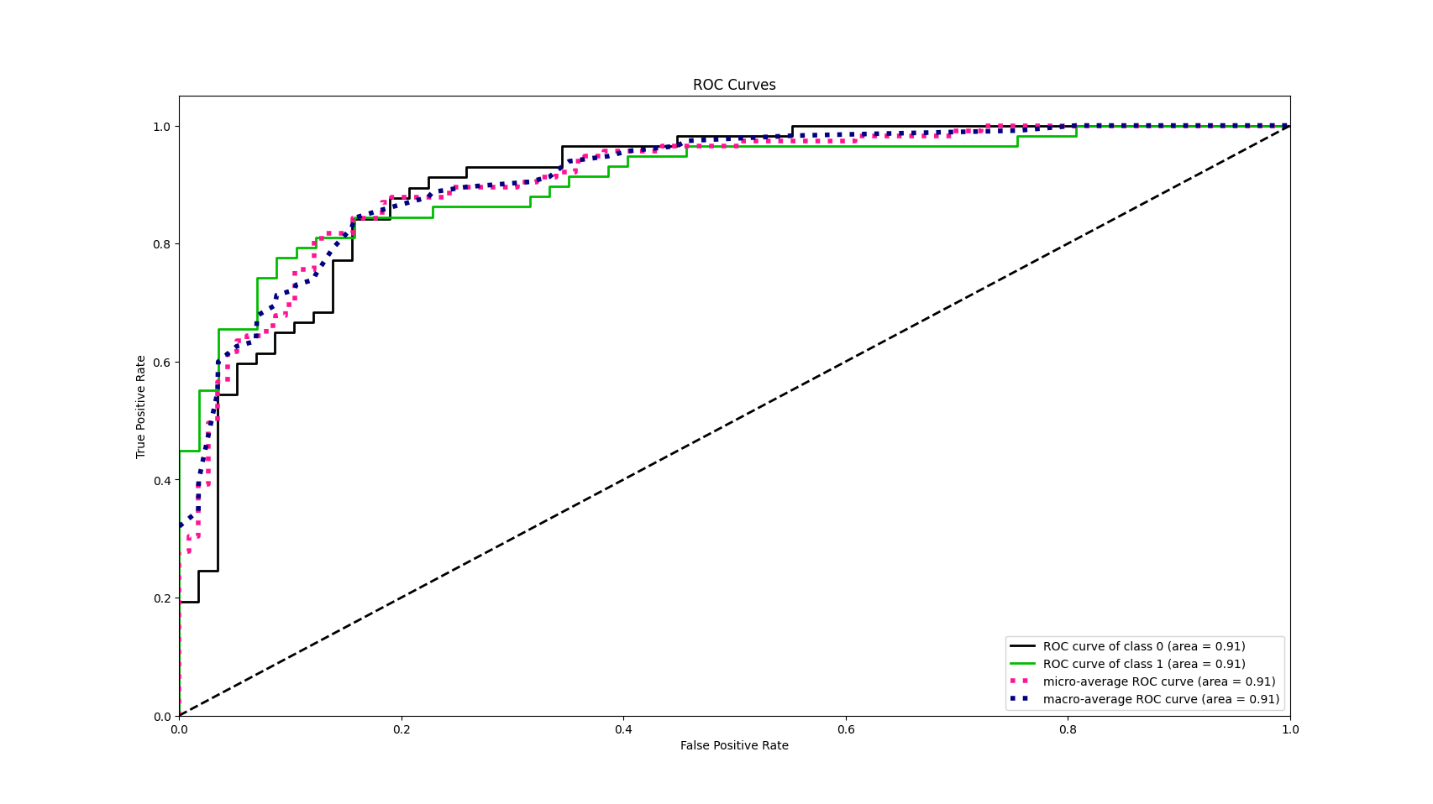


Fig. 15: Random Forest model accuracy based on tree depth.

An average accuracy of 82.% is ultimately where I was unable to improve model performance any further. The model displayed an overall 0.91 area under its ROC curve. Most notably, the model is biased towards predicting Republican, despite there only being a 51/49 split in the data labels. Fig. 16: ROC AUC curve of the Random Forest Classifier (Pedregosa et al, 2011).

Discussion and Improvement

Most notably is the scarcity of data this model works with. With only 459 data series, it is a serious flaw of this model. SMOTE, a synthetic data creation technique, could possibly be used here, however time did not allow for its testing. However, the model does already correctly predict solidly red and blue states. Perhaps more significant for model performance is its ability to determine the outcome of swing states. This is typically done from a county level, but this was not feasible for this project. Finding other variables correlated along party lines will surely help the model’s accuracy. A characteristic of swing states that could indicate their political leaning would also be incredibly beneficial.

Interface

The model is loaded into a StreamLit based “web application” (Streamlit, 2021). It is not hosted online, but can be run and hosted locally. The app requires a number of dependencies, including StreamLit, a number of smaller matplotlib libraries including the Basemap class, and a few other libraies you likely have installed already, such as pandas and numpy. To run the application, open your preferred command prompt window, enter the directory the “PresModel.py” file is located in, and type “streamlit run PresModel.py”. If you do not want to download the extra dependencies, I have also included a couple screenshots. The election year is selected from the drop down menu, and displays the predicted and actual results of the election. The maps that appear, and many of the graphs used in this report, are created with pyplot (Hunter, 2007). The maps in the interface specifically are based on matplotlib’s publicly available code using Basemaps (user “ocefpaf”, 2018).

Image 1, description and maps

A screenshot of a map

Description automatically generated

Image 2, listed states

A screenshot of a map

Description automatically generated

References

Assoc. of Religious Data Archives. 2020. “All Religions - Rankings by Area (States)". <https://www.thearda.com/us-religion/statistics/rankings?u=2&typ=2&cod=9999>

Federal Communications Commission. 2020. “Election results and voting information.” <https://www.fec.gov/introduction-campaign-finance/election-results-and-voting-information/>. Accessed Nov 19, 2023.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. 2020. Array programming with NumPy. Nature 585, 357–362. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). ([Publisher link](https://www.nature.com/articles/s41586-020-2649-2))

J. D. Hunter. 2007. "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95.

Lipka, Michael, Benjamin Wormald. 2016. “How religious is your state?”. <https://www.pewresearch.org/short-reads/2016/02/29/how-religious-is-your-state/?state=alabama>. Accessed Nov 28, 2023.

Ocefpaf. 2018. https://github.com/matplotlib/basemap/blob/master/examples/fillstates.py.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**. Nature Methods, 17(3), 261-272.

Pedregosa et al. 2011. [Scikit-learn: Machine Learning in Python](https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), JMLR 12, pp. 2825-2830.

Pew Research. 2018. “Trends in Party Affiliation Among Demographic Groups”. <https://www.pewresearch.org/politics/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/>. Accessed Nov. 19, 2023.

Small Area Poverty and Income Estimates. 2022. “SAIPE Datasets”. <https://www.census.gov/programs-surveys/saipe/data/datasets.html>. Accessed Nov. 22, 2023.

Streamlit. 2021. <https://streamlit.io/>.

United States. 2020. “2020 Census”. <https://data.census.gov/table/ACSDP1Y2017.DP05?q=United%20States&table=DP05&g=010XX00US&lastDisplayedRow=29&vintage=2017&layer=state&cid=DP05_0001E>. Accessed Nov 15, 2023.

United States. 2010. “State Intercensal Datasets: 2000-2010”. <https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-2000-2010-state.html>. Accessed Nov 18, 2023.

United States. 2010. “State Intercensal Datasets: 1990-2000”. https://www.census.gov/data/datasets/time-series/demo/popest/intercensal-1990-2000-national.html