Project 3 – Voting

Author(s):

Gabriel Jackson

Naad Kundu

Thao Nguyen

Kayla Nguyen

Halbert Nguyen

Ben Willoughby

Omar Zeineddine

DS 3001: Foundations of Machine Learning

May 7, 2024

Summary

The main question we attempted to answer with this project was how well models could predict the outcome of the 2024 presidential election in Virginia. Another area of focus was providing data about how precise our models' predictions were. We developed a detailed plan to execute and analyze our models. After loading and cleaning our CSV-formatted data, we created a linear regression model to predict the total votes and their growth rate for each county in Virginia. We also used polynomial feature expansion, and finally we were able to build a state map which showed each county and what party they were predicted to vote for per our linear regression model.

The results were very interesting: our state map showed which party each county was predicted to vote for, as well as how favored that party was based on a color gradient from dark blue to dark red. Our results showed that more counties in Virginia were predicted to vote red (Republican) than blue (Democratic). Specifically, counties in the South, Southeast, Southwest, and many central Virginian counties were predicted to vote Republican, as well as most counties in Northern Virginia. Many of the blue counties were along the east coast of Virginia (along the Chesapeake Bay), a cluster of counties in the Southwest, and Western/Northwestern Virginia, 2 counties in Central Virginia, and a cluster of counties in Northern Virginia. With many more counties leaning towards red than blue, it seems to paint a picture that Republicans are predicted to win Virginia in the 2024 election, but this model does not paint the full picture, as some counties' populations vary wildly- which could mean that the smaller number of counties that are predicted to vote blue could be enough for a majority to win over Virginia for the democrats.

Finally, we tested the accuracy of our model using R^2, or the coefficient of determination value. Our target variable was candidate votes- and from computing our coefficient of determination, we got a good result of 96.6%, meaning that about 96.6% of variation in candidate votes can be explained by the features we employed in our linear regression model.

Data

The data used for this project came from several provided sources focusing primarily on historical voting data and county level demographic information for Virginia. One of the main datasets used was voting_VA.csv which contained voting data for presidential elections in Virginia from 2000 to 2020 and provided a thorough breakdown of votes for each candidate down to the county voted from. The nhgis_county_data folder contained a wide range of county level summary statistics for every county in the United States, sourced from the IPUMS NHGIS website. The county_adjacencies.csv file contained information on neighboring counties, districts, FIPS county identifiers, and 2022 population estimates for all counties and cities in Virginia. Lastly, a shapefile was provided by the Virginia Geographic Information Network for creating choropleth maps.

During the data cleaning and preparation process, several challenges were encountered. First, the 2020 votes for each candidate were broken into three separate entries: absentee, election day, and provisional votes. These entries had to be combined in order to obtain a total vote count comparable to previous years. Additionally, the given dataset was inconsistent, as voting data was available for some counties in some years but not for all years. This is not something that we were able to easily clean unless we omitted a decent amount of data, so we made sure to account for this aspect when creating our county dummy variables. Furthermore, voting data for other smaller third party candidates such as the Libertarian and Green parties were included, but we decided not to use this data since we predict that these candidates are highly unlikely to win the election. Despite these challenges, the data was successfully cleaned and prepared for analysis. The rich platter of datasets provided to us opens the gate for us to build detailed predictive models and generate insights into voting patterns across Virginia, ultimately predicting the outcome of the 2024 presidential election.

Results

The first learning model that was implemented was a linear regression model. As aforementioned, all of the categorical variables had to be one-hot encoded so the model has numerical inputs to work with. The intended goal with the linear regression model was to predict the total votes for any given candidate based on previous years. After implementing the model, we evaluated the features by extracting their respective coefficients. Because there was a function used to generate interaction terms, the model was able to best understand the data to predict while learning intervariable relationships. In other words, interaction terms appended many of the columns to further examine relations between certain observations of features. It is thus important to note that some variables have multiple candidate names from forming these terms. Among candidates alone, there were nine unique names over the last 20 years, and there are 167 different counties in Virginia to account for.

Table 1. Features of Linear Regression Model and Coefficients.

	Variables	Coefficient
12	AL GORE DONALD J TRUMP	-3.572592e+11
18	AL GORE JOSEPH R BIDEN JR	-2.749271e+11
16	AL GORE JOHN KERRY	-8.774195e+10
21	BARACK OBAMA DONALD J TRUMP	-8.075135e+10
14	AL GORE GEORGE W. BUSH	-7.757942e+10
...	...	...
2	DONALD J TRUMP	7.664185e+10
15	AL GORE HILLARY CLINTON	1.681504e+11
22	BARACK OBAMA DONALD TRUMP	2.003519e+11
17	AL GORE JOHN MCCAIN	2.059024e+11
30	DONALD J TRUMP DONALD TRUMP	2.464118e+11

The model had an R^2 value of 0.9662, which signifies that the model was able to fit and predict the total number of votes relatively well. Specifically, we were able to look at the features and their associated coefficients to investigate the impact of certain features on the predicting ability for total votes

per candidate. At this point in time, however, ‘increasing’ the numerical value of any (encoded) categorical variables is not very insightful to understanding what variables are most impactful on predicting total votes.

To better understand the data that the model predicted, we decided to calculate the average growth rate in the total of numbers per county following every election (Table 2). According to past data, Loudoun, Williamsburg, and King George were the counties with the highest average growth rates respectively, with rates above 20%. Richmond, Roanoke, and Norton were the counties with the lowest average growth rates: in fact, their growth rates were negative, which may imply an overall disinterest in voting as time goes on. From this point on, we used the average growth rates per county to calculate the estimated total number of votes to be received, regardless of party, in the 2024 election. This gave us valuable insight to how the volume of total votes may change in the 2024 presidential election.

Table 2. County Name and Average Growth Rate from 2000-2020.

	County Name	Growth Rate		County Name	Growth Rate
86	LOUDOUN	24.795890	**131**	RICHMOND	-9.670022
160	WILLIAMSBURG	21.416115	**133**	ROANOKE	-3.713910
80	KING GEORGE	20.034121	**110**	NORTON	-1.158739
94	MANASSAS PARK	19.557972	**17**	BUCHANAN	0.513848
103	NEW KENT	19.061035	**41**	DICKENSON	0.530373

After predicting how the number of votes may grow or decrease, we further investigated the growth rates of predicted votes for the Democratic and Republican parties. In order to achieve this, we used the previous linear regression predictive data to compare the breakdowns between parties: specifically, Joseph R. Biden and Donald J. Trump. Both of these candidates were on the 2020 ballot, and thus we filtered the data to only include the number of votes per county from the most recent election. Using the previously-trained model, we extracted the data of votes for candidates Biden and Trump per county. Then, to condense the data into a more comprehensive form, we negated the predicted value of Republican party votes to add to the value of Democratic party votes. As a result, the net candidate votes

were negative if there were more predicted votes for the Republican party than the Democratic party, and vice versa (Table 3).

Table 3. Net Votes per Virginian County.

	County Name	county_fips	Candidate Votes
0	ACCOMACK	51001	-2408.759674
1	ALBEMARLE	51003	14250.927826
2	ALEXANDRIA CITY	51510	20490.802826
3	ALLEGHANY	51005	-5468.275299
4	AMELIA	51007	-5576.212799
...	...	...	...
128	WILLIAMSBURG CITY	51830	-5926.900299
129	WINCHESTER CITY	51840	-4102.478424
130	WISE	51195	-2529.947174
131	WYTHE	51197	-3068.554134
132	YORK	51199	5216.039616

To further visualize the counties and the majority party that was voted for, we implemented a state map that was colored to represent the respective party: red for Republican, and blue for Democratic (Figure 1). We scaled the data of votes using the inverse hyperbolic sine so that stronger, more opaque colored states would highlight the larger difference in party votes. Northern Virginia shows a relatively more diverse mix of colors whereas overall, the state is predominantly red. This geographic visualization, however, does not take into consideration that many counties have drastically smaller populations than others, so upon first glance, concluding that the majority of Virginia voted for the Republican party is premature. However, this geographic visualization may be useful in identifying key areas of interest for future political strategies and campaigns.

Since there are many counties to assess, we also summed all of the values in the ‘Candidate Votes’ column for a value of 450,409.6, which signifies that our linear regression model predicted that there would be about 450,410 more votes in favor of the Democratic candidate over the Republican candidate. This reflects the previously mentioned factor that counties are being disproportionately represented, as it is not accounting for the different populations per county.

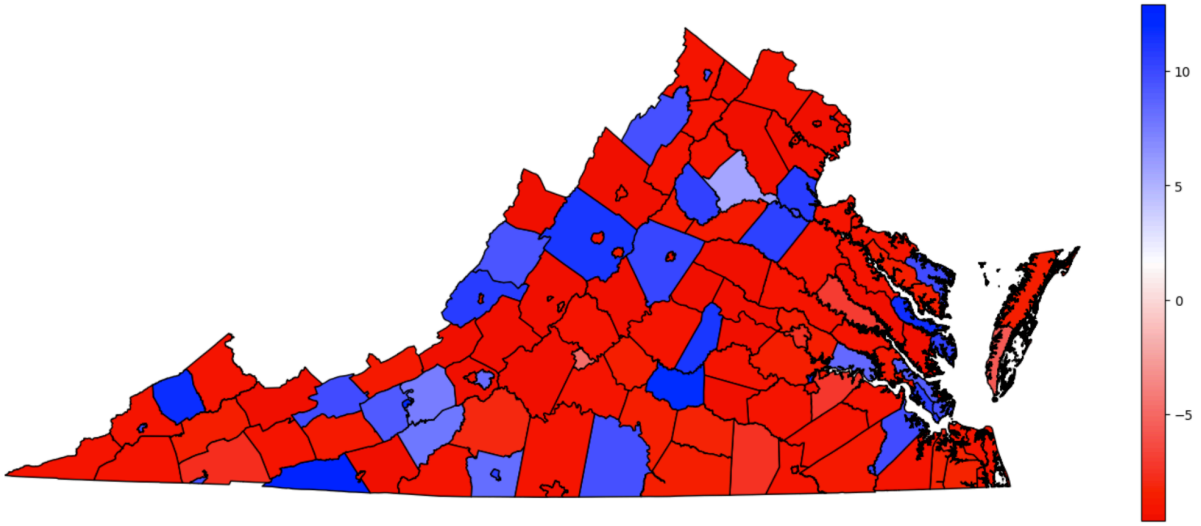

Figure 1. Map of Virginia colored by predicted majority party.

It is also important to note, however, as with any binary classification problem, there is the issue with not representing multiple factors. With many learning models, there are limitations to the predictive abilities from problem to problem. There are many other factors that go into any individual's decision as it pertains to the presidential election, including one's personal beliefs or candidates' specific policies and character. For future implementations, we could further consider associating the party of elected officials for any given county or gathering data from more outside sources to gain a wider perspective on what factors may assist in predicting overall presidential election results.

Conclusion

Overall, our project focused on building models to predict the outcome of the 2024 presidential election in Virginia and provide quantitative information about the precision of the prediction. Through methods of linear regression, polynomial feature expansion, and data visualization via geospatial mapping, we generated predictions for each county in Virginia to see what party they will vote for: Democratic or Republican. We utilized various factors like total votes and which party each county voted for from previous years in order to see how it affected the variable of candidate votes. After thorough data cleaning, we implemented one-hot encoding to convert categorical variables into numerical ones, then we also applied polynomial feature expansion to generate interaction terms up to degree 2. Once we had this information, we built our linear regression model on our target variable, candidate votes. Using our model, we were able to conclude definitive results about growth rate calculation which computed historical growth rates for each county to estimate the total votes for 2024 and also results about the prediction for each candidate by using the trained linear regression model. Finally, we created a state map to visualize the party each county will vote for based on 2024 inputs to the model.

The predictions drawn from the final state map model are pivotal for the opposing party to direct their attention to in an attempt to sway some votes. Focusing on specific regions, the southern counties seem to be predominantly red indicating right-leaning so we predicted them to vote for Trump. The more populated counties, however, like the Chesapeake and northern Virginia counties seem to be mostly blue indicating more left-leaning voters for Biden. Northern Virginia in particular seems to have a mix of colors, reflecting its diverse political landscape, so it will be interesting to see what the actual turnout will be since this area is more densely populated than most other countries meaning it would have a bigger impact in Virginia's overall vote. This geographic visualization helps in identifying key battleground areas and understanding demographic influences on voting behaviors, which could be pivotal for future political strategies and campaigns.

As for criticism, we evaluated the model's performance by computing the coefficient of determination (R^2) to defend our work. The value calculated was 0.9661787960748086, which suggests a

generally good result. Approximately 96.6% of variation in our target variable, candidate votes, is explained by features used in the model. By calculating growth rates for each county to analyze voting trends in order to predict them for the 2024 election, we can rely on past data and results in order to strengthen our own conclusions.

Looking into further exploration and ways to improve our project outside the scope given, we suggested using an inflation metric across the years in order to improve how we forecasted voting behavior over time. This could include adding a socioeconomic factor to further detail voting trends in various demographics within counties using sources like the Federal Reserve Economic Data. Unfortunately, we did not possess the skills to compile all of the data by hand which hindered us from using this data. Our limited knowledge of data scraping in large capacities made collecting the vast information out of scope for this specific project.

In conclusion, while our final model from this project was not perfect, it gave us some insight on voting outcomes for each Virginian county. These outcomes could be influential in the 2024 presidential election as every vote will have an impact. Since this election will be a Biden-Trump rematch, it will be extremely interesting to see how our model using past results from 2020 compares to the final results in the fall. Our hopes with this project and our models will be for others to analyze trends for opposing parties to try to sway voters and focus on these regions.

Appendix

No additional plots or tables were useful to include in this write-up, however the source code can be found within the '.ipynb' files in the project's repository's root at https://github.com/gaboojie/project_voting.