DS 3001 Final Project

Carson Smith, Mackenzie Keeley, Bridget Holt

Fall 2024

Abstract:

This paper investigates the prediction of write-in vote counts in the 2023 Virginia elections using different machine learning methods in order to complete an analysis of the three election datasets. By integrating data from these three data sets - election results, election winners, and election turnout - our study uses a data-driven approach, utilizing linear regression to predict write in votes on both county and state wide elections. Write in votes can often be logistically challenging as they require manual verification and can therefore delay reporting. By predicting write in vote frequency in advance, election officials can better allocate resources to ensure accurate and timely results.

Our analysis employs both a linear regression model and a PCA-based regression model to predict write-in votes at both the state and county levels. However, both of these models exhibited poor predictive performance, with R^2s close to zero (0.0089 and 0.0083, respectively). These outcomes highlight the challenges that severe class imbalances can pose, in this case, where write-in votes represent a very small fraction of the total votes recorded. As well as, highlighting the problems that data ambiguities that complicate the analysis can cause. Explanatory visualizations showed significant variation in write-in vote count from locality to locality and while feature engineering like one-hot encoding and categorical mapping enhanced the data usability it was not totally sufficient to overcome the data usability issues.

In future work there should be a focus on using resampling techniques in order to address the primary issue, class imbalance, as well as exploring other machine learning models that could be better suited to imbalanced data. Despite the limitations, this study still provides interesting insights into write-in voting patterns and underscores the potential for data focused approaches to improve future election planning and predicting.

Introduction:

Write-in votes give voters the ability to support any candidate even if not officially listed on the ballot, enriching democratic participation. However, they can create a logistical challenge for election

officials as they need manual verification in order to ensure validity of the vote. This process requires time and resources, often leading to delays in results being reported. These delays can put a strain of the election workers and reduce the confidence of the public in the election process.

Predicting write-in vote counts at the county level is important for election planning. Forecasts like this allow for election officials to allocate resources, such as equipment and staff, to areas in which high write-in voting is anticipated, helping ensure more time efficient results. This proactive approach improves efficiency and reinforces trust in the outcome of these elections.

Beyond improving election efficiency, this analysis provides politically engaged individuals with a valuable tool to explore potential upsets or unexpected outcomes in elections. increased write-in rates can show voter dissatisfaction with mainstream candidates, offering critical insights for political analysts, strategists, and voters alike. Studying patterns in write-in voting across different localities and office titles allows researchers to uncover trends in voter behavior and electoral dynamics, improving the overall understanding of the political landscape.

In this particular study, we rely on three primary datasets: election turnout, results, and winners from the 2023 Virginia Special Elections. These combined datasets provide a holistic view of voter behavior and election dynamics, enabling our analysis of write-in voter patterns.

The election results dataset holds 66,269 observations across 19 variables, including key features like candidates name, total votes received, party affiliation, write-in vote counts, locality codes, and office titles. This dataset provides the foundation for understanding voting trends at both the county and precinct levels. The turnout dataset holds 2,675 observations across 13 variables that focus primarily on voter participation. Some primary features include, election day turnout, early voting, post-election counts, and mailed absentee voting. It also includes data on active and inactive registered voters which gives us insight into voter engagement for specific reasons. The election winner dataset holds 1,799 observations across 7 variables and records the winning candidates for several different offices across the state. Some key features include the winner's ballot name, office title, and party affiliation. This information is critical for connecting any patterns between write-in votes and election outcomes.

These datasets combine to allow for an integrated analysis of write-in voting trends by blending information on candidates, voter turnout, and election results. They also allow for comparison across different localities and offices, showing the interaction dynamics of participation and voter preferences in Virginia elections.

This study faced a handful of challenges that influenced both the interpretation and analysis of the data. One issue was the ambiguity in write in vote labeling. The write-in votes were labeled, “WRITE IN VOTES”, regardless of the candidate or the voter intent. So while we were still able to collect data on write-in vote frequency we do not have information on the patterns and trends of specific candidates or issues, hindering our ability to have a deeper understanding of why voters chose to write in a name rather than selecting a listed candidate.

A second challenge was merging the three data sets due to inconsistent identifiers. While the datasets contained overlapping information, discrepancies in naming conventions and data structures made direct integration complex. For example, the election turnout dataset used locality names that required alignment with the locality codes in the results dataset, adding an additional layer of data preparation.

During the exploratory phase, key insights emerged that provided a better understanding of the data and informed subsequent modeling efforts. The analysis revealed a substantial class imbalance in the write-in vote counts, with write-in votes representing only a small fraction of the total votes cast. This imbalance posed a challenge for predictive modeling, as the models tended to skew toward predicting low or zero write-in votes, overlooking less frequent but important patterns of high write-in activity.

Visualizations highlighted other notable trends. For example, voter turnout varied significantly across localities and voting methods, with early voting, absentee voting, and election day participation showing distinct patterns. Kernel density plots and bar charts offered insights into the distribution of active and inactive registered voters, providing valuable context for understanding voter engagement at the locality level.

The rest of this paper gives a deeper understanding of the methodology, as we detail the

processing steps, our feature engineering, and our machine learning model used to analyze data. It will then present the results, which include performance metrics like R^2 scores and visual comparisons of predictions. Finally, the paper will conclude with a discussion of the paper's findings, highlighting the limitations encountered, the broader implications of the results, and recommendations for future research to address the challenges of modeling write-in votes.

Data:

Our data consists of election results, turnout, and winners from all Virginia districts in the 2023 November General and Special Elections. The data was sourced from the Virginia Department of Elections website. Key variables include candidates' names, total votes for the candidate, candidate party, locality, office title, total voter turnout, total registered voters, and winner ballot name. In the election results data set, there are in total 19 variables with 66,269 observations. In the election turnout data set, there are in total 13 variables with 2,675 observations. In the election winner data set, there are in total 7 variables with 1,799 observations.

The phenomenon we are studying is the rate of write-in voters per county and if this metric is consistent across office titles. We will use these data sets to predict how many write-in votes will be cast in each county in this coming presidential election.

In using this data, we expect to face challenges in narrowing down the data to the parameters we are investigating. Additionally, the candidate name for each write-in vote is "WRITE IN VOTES." It may become challenging to differentiate between each county and office title as they have the same name for write-in votes. Furthermore, these three data sets can not be directly joined making it more difficult to work with all of them at once. Luckily, there are no commas in the voting data that we will have to clean.

Methods:

Before analyzing the data, it is important to define what an observation is. In our study, each observation is a distinct entry within one of the datasets. In the election results dataset, an observation

captures the voting outcome for a specific candidate in a particular locality and office title. In the turnout dataset, each observation records the total voter turnout within a locality. In the winners dataset, an observation represents the winning candidate for a given race, location, and office title. Altogether, these observations provide detailed information on election results, voter participation, and race outcomes across locations and offices.

When planning our analysis, we decided it would align best with a supervised learning model as we are relying on election data from the past to predict future election outcomes. Concurrently, this is best suited for a regression-based analysis, as it involves predicting a continuous variable: the count of write-in votes.

To predict write-in votes, we plan to start with a linear regression model on an 80/20 test set to establish an understanding of the current voting trends. We will also do a locality-based analysis to understand which candidates are predicted to win in each county. We will largely predict this using individual voters' registered parties and their locality in a linear regression model

We will know if our approach works by comparing how our analysis of data from the "Virginia Election Turnout" and "Virginia Election Results" datasets compare with each other, as well as the "Virginia Election Winners" dataset. As we are looking to predict winning candidates both in the election overall and on a locality-based level (county), success will be measured in terms of the percentage of our write-in data that comes within a $\pm$ 10 range of the comparable data.

There were three main weaknesses we anticipated being an issue when analyzing our data. Principally, difficulty in differentiating "WRITE IN VOTES" entries. These votes had the ability to cause ambiguity in the data. This can make it hard to track any pattern that may arise in the voters and changes the learning curve of the model as it counts any write-in voters as their own single category. In order to address this we may need to add/ rely on additional information like office title or locality to distinguish these voters. Second, we anticipated facing data integration issues. These data sets can not be conveniently joined, which can make it difficult to merge certain features across datasets. We may need to create a unique variable to merge on or merge on the closest matching variables (e.g. locality).

Lastly, we anticipated facing class imbalance challenges. In this data, write-in votes are much rarer than votes for named candidates, creating an imbalanced dataset. This imbalance can be problematic as it leads to the model being biased toward the majority class—predicting low or zero write-in counts more frequently. As a result, the model may overlook patterns that could help identify when write-in counts are higher, reducing its ability to accurately capture those less common but important instances. Managing this imbalance is important to ensure the model remains sensitive to all outcomes and doesn't excessively favor the more frequent class. In order to address this, we may need to use resampling techniques or metrics that account for imbalance, like F1 scores.

If our approach doesn't succeed, it could reveal which features of our data were insufficient in capturing the nuances of write-in voter behavior. We might learn that a different model could have been better suited for this dataset, that additional or more specific data was needed to improve prediction, or that adjustments to the model are necessary to address issues like class imbalance or data ambiguity in the "WRITE IN VOTES" entries. Understanding this would assist us in improving our analysis, helping us make informed choices about feature adjustments and model selection, and potentially exploring other methods to better predict write-in vote trends across counties and office titles.

Results:

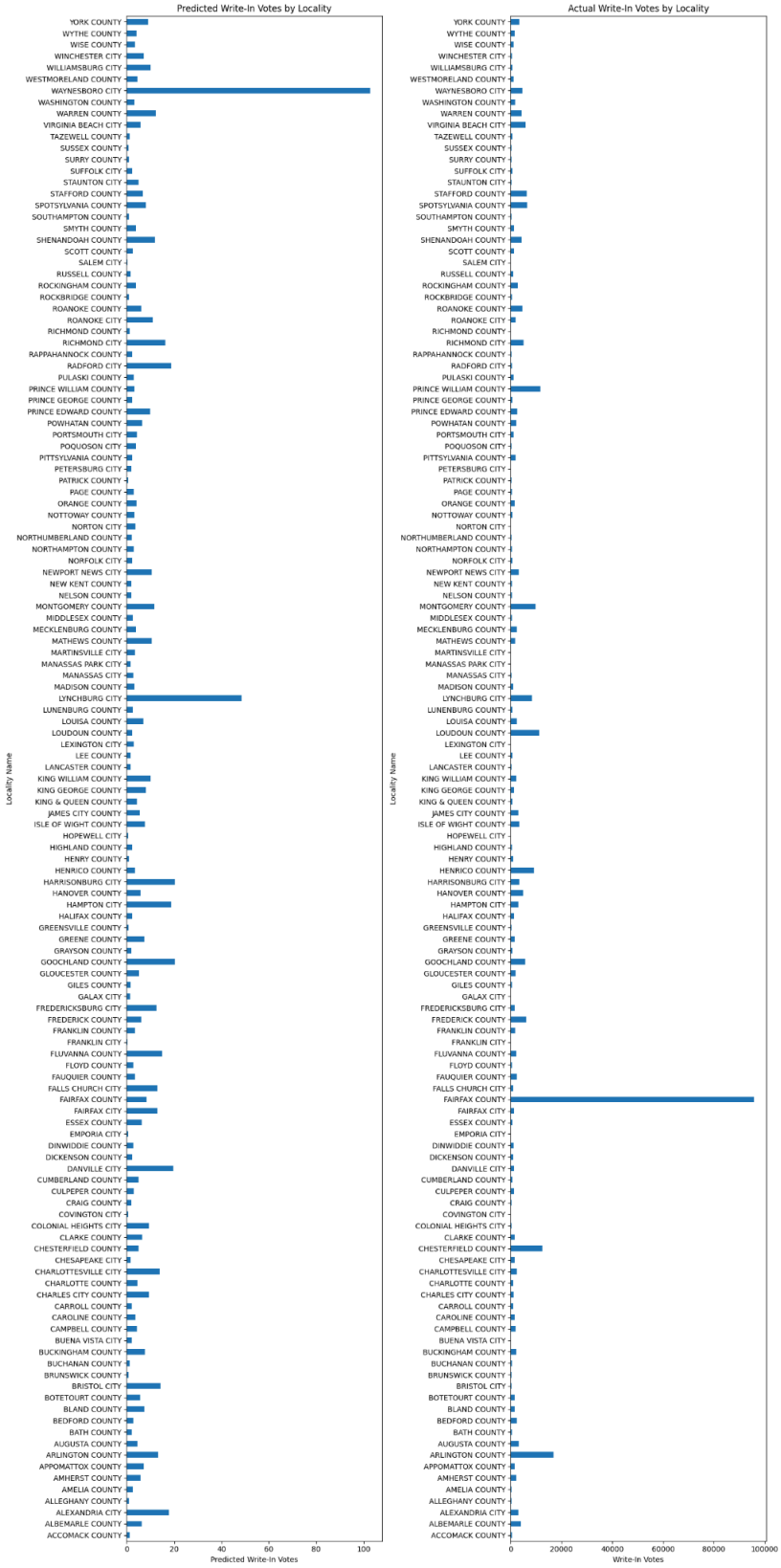
The main aim of our analysis was to predict the number of write-in votes by locality for the 2024 presidential election. However, our regression model faced challenges in reaching a strong predictive performance due to the significant class imbalance, along with a handful of other issues. Our results, which include mean, R^2, and mean squared errors (MSE), give us a better understanding of these challenges.

After doing a linear regression model on an 80/20 test set to establish an understanding of the current voting trends, we used this data to predict the number of write-in votes per locality. Even just by looking at the table, it was clear the predictions were largely incorrect. We ran an R^2 and MSE on the data to understand the proportion of variance that was explained and how close the model's predictions

were to the winners dataset, respectfully. We discovered an R^2 of 0.008893355530268043 and an MSE of 677.4810505305412. This minimal R^2 value tells us that the model is not a good fit for the data and does not explain nor accurately predict the variations in the data. The large MSE value tells us that the predicted values are dispersed widely around the data's mean and the model exhibits poor predictive performance. The poor performance of the model can be seen below in **Figure 1.**

After running a PCA-based model, we discovered similar numbers. The code output a R^2 value of 0.008280913407647694 and an MSE of 856.6816875923196. These results suggest that very little of the variance in write-in vote counts is explained by the model, indicating a weak linear relationship between the feature and target variable.

Figure 1


Predicted Write-In Votes by Locality
Actual Write-In Votes by Locality
Locality Name
Predicted Write-In Votes
Write-In Votes
0
20
40
60
80
100
0
20000
40000
60000
80000
100000

Conclusion:

Predicting write-in votes from the 2023 Virginia election data proved to be a complex task, with significant challenges stemming from class imbalance, limited feature sets, and the choice of dimensionality reduction and modeling techniques. While Principal Component Analysis (PCA) was used as a dimensionality reduction method, its limitations contributed to the task's overall lack of success.

One of the most critical issues with the dataset was the clear class imbalance. Write-in votes are rare compared to votes for campaigning candidates, resulting in a dataset where instances with high write-in vote counts were disproportionately underrepresented. This imbalance led the model to bias predictions toward the majority class, making it effective at predicting low or zero write-in votes but unable to accurately identify significant write-in votes. Addressing this imbalance is key to improving prediction accuracy. Techniques such as oversampling the minority class, undersampling the majority class, or implementing weighted loss functions could help lessen the bias toward the majority class. Without these interventions, the model is struggling to learn the characteristics of high write-in vote instances, contributing to significant prediction errors.

Further, the results of the linear regression and PCA-based models revealed low R-squared values, suggesting a weak linear relationship between the chosen features, party and locality, and the target variable, write-in votes. This outcome highlights a limitation: the assumption of linearity in PCA and linear regression was not suitable for capturing the non-linear patterns that likely influence write-in voting behavior. To address this, alternative modeling approaches that can accommodate non-linear relationships, such as decision trees, random forests, or neural networks, should be explored. These methods are better at uncovering complex interactions between features and the target variable, usually leading to improved predictive performance.

The feature set used in the analysis was also restricted, including only ‘Party_Float’ and ‘LocalityCode’ as predictors. This limitation constrained the model’s ability to capture the full range of factors influencing write-in votes. Write-in voting behavior is likely influenced by a variety of additional factors, such as voter demographics (age, education level, income, etc), local politics, historical voting

patterns, and candidate popularity. Incorporating external data sources, such as voter surveys, could enhance the model's predictive capabilities.

While PCA is a powerful tool for dimensionality reduction, it assumes that the most important information lies in the directions of greatest variance. This assumption may not hold true in this context, where the variance captured by the principal components does not necessarily correlate with the factors driving write-in votes. Additionally, PCA is a linear technique, making it less effective at capturing non-linear relationships in the data.

References

Election Results. https://enr.elections.virginia.gov/results/public/Virginia/elections/2023-Nov-Gen/reports. Accessed 14 Dec. 2024.