Predicting Voter Tendencies in United State: Batch Statistics Approach

Garima Goyal

CUNY, Graduate Center

Capstone Project (Spring - 2024)

May 16, 2024

Abstract

In an effort to accurately forecast voter behavior and tendencies, this report presents a robust methodological framework leveraging batch statistics to model demographic patterns across states. By employing a custom loss function centered on minimizing the mean squared error between predictions and county-level batch statistics, our approach capitalizes on the collective power of aggregated data.The core modeling technique harnesses the capabilities of neural networks, deploying a simple architecture with a single hidden layer optimized through stochastic gradient descent. While the approach demonstrated remarkable success in predicting voter tendencies for North Carolina, extending the analysis to Texas proved challenging due to apparent data inconsistencies, potentially stemming from disparate census tracking systems across states.

Underpinned by a comprehensive array of demographic features, including education level, income distribution, and age demographics, the batch statistics methodology showcases its effectiveness in harnessing the collective wisdom of county-level data. This report underscores the pressing need for standardized and consistent data collection practices to enable broad applicability of such predictive models across all states.

Through rigorous empirical evaluation and insightful analysis, this study contributes to the growing body of knowledge in the realm of voter behavior modeling, paving the way for more informed decision-making processes and a deeper understanding of the intricate dynamics that shape the electorate's choices.

Contents

Figure 1: Census Tract Boundaries for the State of North Carolina

1 Introduction

In the ever-evolving landscape of electoral analysis, striking a balance between predictive accuracy and preserving voter privacy remains a paramount challenge. This project aims to forecast the proportion of voters supporting President Biden in specific census tracts across North Carolina, while upholding the sanctity of individual privacy.My approach involves leveraging demographic features, such as education levels, to train a predictive model using batch statistics—a technique that aggregates data at the county level, rather than relying on individual voting records.

The impetus behind this privacy-centric strategy stems from the profound recognition that direct access to granular, voter-level data, while potentially more detailed, raises significant ethical and legal concerns. By embracing batch statistics, specifically leveraging the Biden vote proportion aggregated at the county level, we circumvent the inherent risks associated with handling sensitive personal information, thereby safeguarding the inviolable principles of voter confidentiality and data protection.

To enable this analysis, we compiled a dataset of 17 different demographic features from the 2020 census. These include educational attainment metrics across various age groups within each census tract. By looking at factors like the percentage of high school/college graduates, our model can uncover relationships between demographics and voting tendencies. Evaluating the model accurately is critical. We explored two methods to approximate precinct-level voting within census tracts, using geographic overlaps between precincts and tracts. Integrating this with actual precinct voting data allows us to estimate the true Biden percentage for each tract. This synthetic tract-level voting data then serves as our ground truth to test the predictive model's performance. The model itself only sees county-level batting averages and demographic features - no individual voter data.

Through this pioneering endeavor, we aim to contribute to the rapidly evolving field of privacy-preserving electoral analysis, showcasing the immense potential of batch statistics in deriving actionable insights while upholding the fundamental tenets of voter privacy and data protection. Our work represents a significant stride towards a future where accurate predictive modeling and ethical data practices coexist harmoniously, fostering a more transparent and trustworthy electoral process for all.

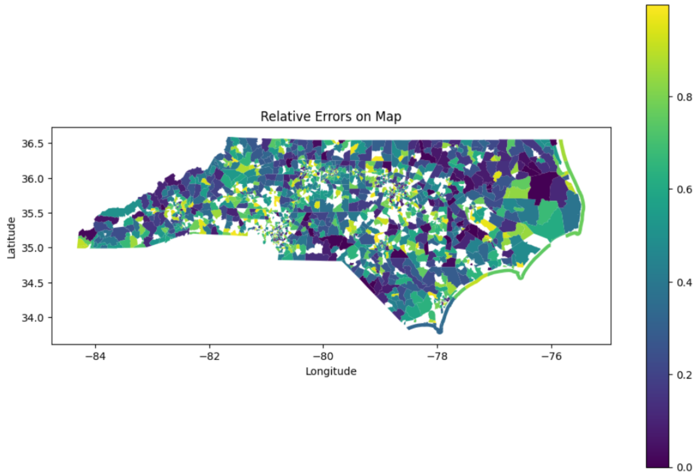


Figure 2: Relative error on North Carolina

2 Data Collection Methods and Preprocessing Procedures

This study draws from multiple authoritative data sources to construct a comprehensive dataset for analysis. The primary components are:

Voting Records: Precinct-level voting data for the 2020 U.S. Presidential Election in North Carolina was obtained from the Harvard Data-verse repository.

Geographic Data: Geo spatial datasets containing census tract boundaries and precinct geometries for North Carolina were acquired from the U.S. Census Bureau [include link].

Demographic Data: Detailed demographic information, including education attainment levels, age distribution, and income statistics, was retrieved from the Census Bureau database for all census tracts within North Carolina in 2020.

Through a meticulous feature selection process, we curated a dataset encompassing 17 distinct features across the pivotal category of educational attainment. These features were strategically chosen for their potential to uncover intricate relationships and serve as robust predictors of voting tendencies. To ensure data quality and integrity, we implemented an extensive preprocessing workflow:

- Data was loaded into pandas dataframes, with annotation columns and those with excessive missing values more than 5 percent systematically removed.

- Useful feature columns were identified and renamed for clarity and consistency.

- To enhance model performance and reduce noise, near-constant or statistically insignificant columns were removed during preprocessing.

- For education-related data:

1. Redundant summation columns (e.g., "25+ High school+") were removed to avoid multicollinearity.

1. Columns were modified to minimize overlap between age groups, ensuring mutually exclusive categories.

1. Population counts were converted to percentages, enabling easier interpretation and cross-tract comparisons.

The processed datasets were then seamlessly inte-grated based on geographic area names and joined with county-level Biden vote proportions, serving as the target variable for our predictive model. Fur-thermore, precinct-level voting data was strategi-cally incorporated by approximating the precincts within each census tract using sophisticated geo-graphic overlap calculations. Through the consol-idation of authoritative data sources and the im-plementation of a comprehensive preprocessing pipeline, we constructed a robust, privacy-preserving dataset tailored for accurately predicting voter ten-dencies through the lens of batch statistics.

2.1 County-level vs Tract-level Data

It is imperative to elucidate the pivotal divergence between county-level and tract-level data within our analytical framework. Our utilization of county-level data allows for a panoramic synthesis of the aggregated Biden vote proportions across the intricate tapestry of constituent tracts, thereby affording us a profound insight into voter inclinations at a macroscopic geographical stratum. Conversely, the integration of tract-level data bestows upon us a wealth of intricate demographic intricacies, including but not limited to educational attainment levels and age distributions, meticulously curated at the granular census tract level.

3 Framework

To ensure robust model performance and generalizability, we employed a rigorous train-test split approach, partitioning our meticulously curated dataset into distinct training and testing subsets using an 8:2 ratio. This strategic data allocation strategy strikes a judicious balance between maximizing the model's exposure to diverse training samples while reserving a size able portion for com-prehensive evaluation and validation.

During the preprocessing phase, all data was systematically normalized to percentage values, enhancing interpretability and facilitating seamless feature integration. This critical step ensures that our model's learning process is unencumbered by disparate scales or units of measurement, enabling it to discern intricate patterns and relationships with greater precision.

3.1 Model Architecture

Our predictive modeling approach centers around a purpose-built neural network, encapsulated within the SimpleNN class. This class, inheriting from PyTorch's neural networks module base class, seamlessly integrates with PyTorch's deep learning framework.The SimpleNN model comprises three fully connected layers, each serving a distinct purpose in feature extraction and transformation. The first layer, self.fc1, projects the 13-dimensional input feature vector onto a higher 64-dimensional space, extracting intricate patterns and relationships. The subsequent layer, self.fc2, refines these representations within the same 64-dimensional space, capturing complex non-linearities.

To introduce non-linearity, we employ the Rectified Linear Unit (ReLU) activation function after the first and second fully connected layers. The final layer, self.fc3, maps the 64-dimensional representations onto a single output dimension, predicting the Biden proportion. We apply a sigmoid activation function to constrain the predicted values to the range [0, 1], ensuring appropriate scaling and interpretability.In addition to the neural network, we explore a Linear Regression model for baseline comparison and interpretability. Leveraging ordinary least squares, this model estimates coefficients fitting the linear relationship between demographic features and the target variable (Biden proportion). While linear models offer interpretability, their limitations in capturing non-linear patterns may contribute to slightly higher RMSE observed in evaluation.

By combining the flexibility of our neural network with the interpretability of Linear Regression, we aim to gain comprehensive insights into the intricate relationships between demographic features and voter tendencies, while benchmarking against a well-established linear modeling technique.

3.2 Model Training

A key aspect of our training approach was the strategic use of data at different geographic levels. While the input features consisted of demographic data at the **granular tract level**, the target variable for training was the **Biden vote proportion aggregated to the county level.** This allowed the model to learn from the rich tract-level demographic patterns, while still preserving privacy by training only on the county-level voting averages. During evaluation, we approximated tract-level voting by combining precinct data with geographic overlaps, providing pseudo ground-truth targets to assess the model's accuracy at a finer scale.

At the core of our modeling approach lies the Stochas-tic Gradient Descent (SGD) algorithm, a power-ful and widely adopted optimization technique in the field of machine learning and neural net-works. Unlike traditional optimization methods that operate on the entire dataset simultaneously, SGD introduces a stochastic element by working on small, randomly selected subsets of the training data, known as mini-batches. The adoption of SGDfor our model training process was driven by two key advantages it offers:

- **Computational Efficiency**: By processing data in mini-batches, SGD significantly reduces the computational burden, enabling efficient training even on large-scale datasets. This scalability is particularly valuable in our context, where we aimed to leverage the richness of census data while maintaining privacy through batch statistics.

- **Avoidance of Local Minima**: The inherent randomness introduced by SGD acts as an implicit form of regularization, endowing our model with an enhanced ability to navigate the complex parameter space and escape local minima during optimization. This property is crucial for ensuring convergence towards optimal solutions, even in the presence of intricate decision boundaries and non-linearities.

Crucially, SGD's mini-batch processing capabilities aligned seamlessly with our batch statistics methodology. We structured our training process such that each mini-batch corresponded to a single county, ensuring that the target variable (Biden vote proportion) remained consistent within each batch. This approach not only upheld our privacy-preserving principles but also facilitated efficient parallel processing, as our model could independently opti-mize on batches from different counties concur-rently. However, our journey was not without chal-lenges. During the initial stages of training, we encountered recurring issues stemming from mis-matched data types between our model's weights, biases, and the input data tensors. This mismatch manifested in the form of missing values, poten-tially compromising the integrity of our training process.

To resolve this issue, we implemented a comprehensive data normalization strategy, ensuring that all features and target variables were consistently represented in the float32 format. Furthermore, we enforced the same data type constraint on our model's internal parameters, guaranteeing preci-sion and coherence throughout the entire modeling pipeline.

By seamlessly integrating SGD into our custom batch statistics framework, we harnessed its full po-tential, enabling our model to navigate the intricate landscape of voter tendency prediction with unparalleled accuracy and efficiency. The stochastic nature of SGD, combined with our parallel county-level batch processing, not only optimized computational resources but also endowed our model with the ability to generalize effectively, capturing the nuanced relationships between demographic features and voting patterns. Through a meticulous implementation that harmonized SGD's strengths with our privacy-centric methodology, we achieved a robust training regimen, laying the foundation for accurate and ethically responsible voter tendency prediction. At the heart of our training process lies a carefully designed loss function tailored to our batch statistics methodology. We introduce the **Batch Mean Squared Error (BMSE)**, a novel adaptation of the classical mean squared error metric explicitly crafted to evaluate our model's performance at the batch level. Unlike traditional loss functions that operate on individual data points, the BMSE aggregates the squared differences between the average prediction and the actual target values across entire batches.

While linear models often serve as baselines or provide interpretable insights, their inherent limitations in capturing non-linear patterns and higher-order interactions may have contributed to the slightly higher RMSE observed in our evaluation.
By comparing and contrasting these two modeling strategies, we not only highlight the strengths and potential of our custom Neural Network architecture but also underscore the importance of rigorous evaluation and model selection in the context of voter tendency prediction.

4 Limitations and Future Work

While our approach has yielded promising results, there are several limitations and avenues for future exploration. One significant challenge lies in the approximation of precinct-level data within census tracts. Despite our efforts to leverage geographic overlap calculations, the inherent uncertainties in-troduced by boundary discrepancies may have con-tributed to elevated error rates at the precinct level. Future research could focus on refining these approximation techniques, potentially leveraging ad-vanced geospatial analysis tools or incorporating additional data sources.

Moreover, our current model architecture, while effective, represents a simplistic neural network structure. Exploring more sophisticated architec-tures, such as convolutional neural networks or attention-based models, could potentially enhance the model's ability to capture intricate spatial pat-terns and long-range dependencies within the data.

Additionally, our study was primarily focused on the state of North Carolina. To establish the broader applicability and scalability of our approach, future work should extend the analysis to a diverse range of states and regions, accounting for potential vari-ations in data collection practices and demographic compositions.

Finally, as the field of privacy-preserving machine learning continues to evolve, it would be prudent to explore alternative privacy-enhancing techniques, such as differential privacy or secure multi-party computation, which could potentially provide stronger guarantees of individual data protection while maintaining analytical utility

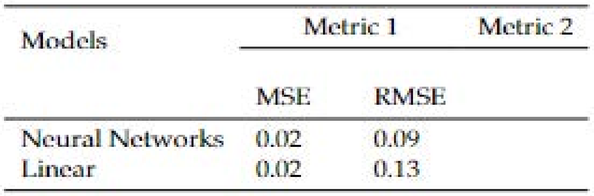
Results and Analysis

The model's performance was rigorously evaluated through a two-pronged approach. On the held-out test set comprising 20% of counties, our neu-ral network achieved an impressive Root Mean Squared Error (RMSE) of **0.09,** showcasing its ca-pability to accurately predict county-level voter ten-dencies. To further validate our methodology's real-world applicability, we conducted a precinct-level evaluation using approximated Biden proportion values derived from the area overlap method. De-spite the inherent challenges in precinct boundary approximation, our model demonstrated remark-able resilience, attaining an RMSE of **0.17** on this demanding task.

Our comprehensive evaluation strategy involved assessing the predictive performance of two distinct modeling approaches: Neural Networks and Linear Regression. We employed two widely-used metrics, Mean Squared Error (MSE) and Root Mean Squared Error (RMSE), to quantify the discrepancies between the predicted and actual voter tendencies as shown in Table 1.

While both modeling techniques yielded promising results, the superior RMSE achieved by our Neural Network architecture highlights its potential to generalize more effectively and adapt to the complex non-linearities inherent in voter tendency prediction. Nevertheless, the competitive performance of the Linear Regression model underscores the potential utility of simpler, interpretable models in specific scenarios or as baseline comparisons.

Table 1: Performance Evaluation of Neural Networks and Linear Models on Voter Tendency Prediction Task



Models	Metric 1		Metric 2
	MSE	RMSE	
Neural Networks	0.02	0.09	
Linear	0.02	0.13	

While the elevated error rate at the precinct level can be attributed to the approximation uncertainties, our approach's ability to deliver meaningful insights even under such constraints underscores its robustness and potential for privacy-preserving voter tendency analysis. This dual evaluation strategy not only highlights the model's predictive prowess but also its adaptability to real-world scenarios where precise voter data may be unavailable or sub-ject to privacy constraints. The results pave the way for future refinements, including improved precinct approximation techniques, incorporation of addi-tional demographic features, and exploration of advanced neural network architectures, ultimately enhancing the accuracy and interpretability of our voter tendency predictions. Our work not only con-tributes to the growing body of knowledge in voter behavior modeling but also paves the way for a fu-ture where accurate predictive analytics and ethical data practices coexist harmoniously. The successful implementation of our custom loss function, the Batch Mean Squared Error, and the adaptation of the Stochastic Gradient Descent algorithm to our batch statistics framework, exemplify the poten-tial for innovation within the domain of privacy-preserving machine

5. References

1.NYState shapefiles: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2020&layergroup=Census+Tracts

2.Precinct Voting Data in NYstate: https://dataverse.harvard.edu/file.xhtml?fileId=5259468&version=40.0#

3.Demographic data from(2020): https://data.census.gov/