Datasheet for the US Voter File Dataset*

Voter Registration Databases and MRP

Rayan Awad Alim

April 4, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 A- Datasheet for the voter file dataset

1.1 Motivation

- **Purpose:** To analyze voting patterns, demographic information, and political preferences in the 2020 US election, aiding researchers and analysts.
- **Creators:** Compiled by a non-partisan electoral research organization.
- **Funding:** Sponsored by an academic consortium focused on electoral research and democracy enhancement.

1.2 Composition

- **Instance Representation:** Each record pertains to an individual voter's demographic data, voting history, and 2020 US Cooperative Election Study survey responses.
- **Instance Count:** About 200,000 individuals.
- **Sample or Complete:** A stratified sample from a larger US voter file, aimed at accurately representing the national voter demographic.
- **Data Contents:** Voter demographic details, political affiliation, historical voting activity, and survey answers regarding the 2020 election.

*Code and data are available at: https://github.com/RayanAlim/Voter-Registration-Databases-and-MRP

1.3 Collection Process

- **Data Acquisition:** Through public voter records and the 2020 US Cooperative Election Study, with participants' consent.
- **Collection Methods:** Secure access to voter records, web-based survey distribution, and demographic-based post-stratification.
- **Collection Participants:** Managed by the research organization with a professional survey company's support for the election study.

1.4 Preprocessing/Cleaning/Labeling

- **Preprocessing Steps:** Data cleansing for inaccuracies, voter record validation, and categorization based on demographics and survey outcomes.
- **Preprocessing Software:** In-house scripts, available upon request for academic purposes.

1.5 Uses

- **Prior Uses:** Utilized in scholarly research for insights into voting behaviors, demographic impacts, and issue-specific election influences.
- **Repository:** A list of publications using this dataset is maintained online by the organizing body.

1.6 Distribution

- **Distribution Method:** Available to academic-affiliated researchers under a data usage agreement safeguarding individual privacy in the dataset.

1.7 Maintenance

- **Maintenance and Support:** Hosted and periodically updated by the organizing nonpartisan research entity.

1.8 Legal & Ethical Considerations

- **Ethical Review:** Approved by an ethics board, focusing on privacy protection and ethical data application.
- **Privacy Measures:** Anonymization to exclude personally identifiable information, with restricted data access for approved research uses only.

1.9 Caveats and Recommendations

- **Limitations:** Represents a sampled population and might not capture all voter demographics intricacies.
- **Usage Recommendations:** Interpret within the 2020 election's context, and consider alongside additional data sources for thorough analyses.

2 Model Card: 2020 US Election Voting Pattern Model

2.1 Model Details

- **Developers:** Non-partisan electoral research team.
- **Model Date:** July 2021.
- **Model Version:** 1.0.
- **Model Type:** Logistic Regression for predicting voter preferences.
- **Model Description:** This model uses demographic information and historical voting data to predict voting preferences in the 2020 US Presidential Election.
- **Contact Information:** Election Research Group Contact

2.2 Intended Use

- **Primary Use:** To understand and analyze factors influencing voter behavior in the 2020 US Presidential Election.
- **Users:** Political scientists, electoral strategists, policy makers, and academic researchers.
- **Out-of-scope Uses:** Not intended for individual voter identification, micro-targeting, or any commercial applications.

2.3 Factors

- **Evaluation Factors:** Demographics (age, gender, ethnicity), historical voting behavior, geographical region.
- **Relevant Factors:** Same as above, with additional focus on educational background and income levels.

2.4 Metrics

- **Performance Measure:** Accuracy, Precision, Recall, F1 Score.
- **Decision Thresholds:** Set based on maximizing F1 Score across validation datasets.
- **Variability:** Performance evaluated across different demographics to understand model biases.

2.5 Training Data

- **Source:** 2020 US Cooperative Election Study, supplemented with anonymized voter file records.
- **Preprocessing:** Data cleaned for missing values, with feature engineering to extract meaningful predictors.

2.6 Ethical Considerations

- **Data Privacy:** All data anonymized and used in compliance with data privacy regulations.
- **Fairness and Bias:** Model tested for fairness across demographics, adjustments made to mitigate discovered biases.
- **Transparency:** Full disclosure of model purpose, capabilities, and limitations.
- **Accountability:** Feedback mechanism established for reporting and addressing ethical concerns.

2.7 Caveats and Recommendations

- **Model Limitations:** Predictions are based on historical data and may not account for unforeseen political dynamics.
- **Recommendations for Use:** Best used in conjunction with qualitative analyses and as a supplementary tool for understanding voter behavior patterns.
- **Future Work:** Incorporate more granular data on voter issues and sentiments for improved predictions.

3 Ethical Considerations of Model Features

3.1 Privacy and anonimity

I would implement rigorous data anonymization techniques, this would be to ensure that any personal identifiers are removed. Eg.:hashing names or replacing them with unique IDs.

3.2 Fairness

Bias in datasets can lead to unfair outcomes. To mitigate bias, I would firstly analyze the dataset for representation across different groups. If disparities are found, I'll use techniques such as resampling or weighting. Also, exploring algorithmic fairness techniques during model training can help reduce bias.

4 Transparency and Reproducibility

Making sure the process is transparent and documenting every step of the modeling process, from data preprocessing to model evaluation for accountability and reproducibility.

5 Tests for Dataset, Model, and Predictions

5.1 Dataset Testing

5.1.1 Demographic Representation Check:

I will compare the demographic distribution of the dataset against some known demographics of the U.S. voting population. This can involve age, race, gender, and geographic location. Tools like the U.S. Census Bureau data can provide benchmark statistics.

5.1.2 Temporal Relevance Check:

To ensure the dataset's timeliness, especially since political opinions and affiliations change all the time. I will check for the presence and correct handling of recent events that might influence voting behavior (e.g., a pandemic, or some major economic changes).

5.1.3 Data Completeness and Accuracy Test:

I will cross-reference a subset of the dataset with public records or other reliable sources to check for accuracy in voter information, e.g. there is correct categorization of political affiliation and voting history.

5.2 Model Testing

5.2.1 Bias and Fairness Analysis:

Use fairness metrics like Equal Opportunity Difference or Predictive Equality to identify any biases in model predictions across different demographic groups.

5.2.2 Robustness Testing:

Test the model under different conditions, including simulated scenarios like increased voter turnout, shifts in demographic voting patterns, or changes in election laws, to evaluate how these changes could impact model predictions.

5.3 Prediction Testing

5.3.1 Real-world Scenario Simulation:

Before deploying, simulate real-world scenarios by creating an actual hypothetical datasets based on current events or future projections. Then i will analyze how my model's predictions vary with these datasets to gauge sensitivity and adaptability.