

Datasheet for CES 2020 Voter Behaviour Prediction Model

Alexander Sun

April 3, 2024

1. Motivation

- **Why was the dataset created?** The dataset was created to support research and analysis on voting patterns, electoral engagement, and demographic representation in electoral processes.
- **Who created the dataset?** This dataset was compiled by governmental election bodies and political organizations.
- **Who funded the creation of the dataset?** Funding sources could include governmental grants, academic institutions, and private investors.

2. Composition

- **What do the instances represent?** Each instance in the dataset represents an individual voter, including their voting history, demographic information, and potentially their party affiliation.
- **How many instances are there in total?** There are approximately 100 million instances
- **What data does each instance consist of?** Typically, an instance would include the voter's ID, age, gender, race/ethnicity, address, party affiliation, and voting history (e.g., elections participated in).

3. Collection Process

- **How was the data collected?** Data is collected from official voter registration records, which are maintained by local, state, and national election offices.
- **Who was responsible for data collection?** Data collection is overseen by the government
- **Were any ethical considerations taken into account?** Ethical considerations include ensuring voter privacy, adhering to legal restrictions on voter data usage, and anonymizing data to prevent individual identification.

4. Cleaning

- **Was any preprocessing/cleaning/labeling of the data done?** Anonymizing personal identifiers, correcting erroneous entries, and standardizing data formats.
- **Was the raw data saved in addition to the preprocessed/cleaned data?** Yes
- **Is the software used for preprocessing/cleaning/labeling available?** Yes

5. Uses

- **Has the dataset been used for any tasks already?** Tasks include studies on voter turnout, analyses of demographic influences on voting behavior, and models predicting future election outcomes.

6. Distribution

- **How is the dataset distributed?** The dataset is distributed via the government database
- **Is there a cost associated with the dataset?** Information from the dataset is freely available
- **Are there any legal or ethical issues with dataset distribution?** Legal and ethical considerations include data privacy laws, consent for data use, and restrictions on the redistribution of voter information.

7. Maintenance

- **How is the dataset maintained?** The data is maintained by the government and updated every election.