Datasheet for 'Voter File Dataset'*

Mohammed Yusuf Shaikh

2024-04-01

1 Voter Dataset

1.1 Motivation:

What was the main point in creating the dataset? The dataset was designed to identify the reasons for vote, voters' age and gender, as well as what trends the elections in 2020 had. The funding of the construction of such dataset should be identified and hence fundamental rights, especially privacy of people are not at stake. Filmed in a fictional tie-breaker scenario, so let's imagine it was funded by non-partisan (transparent and objective) research organization that is interested in the role of an individual on the electoral process.

1.2 Composition:

Which particular cases are go signature in the collection? Every entry is about the US voter, namely, that voter's personal information (Demographics), voter history, and surveys response.bHow real or true is it? What is so real or true about it . This is contingent on the actual data, but take the case in which might get to about a thousand and a half to give a precise representation of the voters.

1.3 Collection Process:

What is the way to reveal the data from this work? Datasets were drawn from multiple sources of the elections in the midterm Congressional elections: a combination of voter files, pre- and post- election surveys, and the Cooperative Elections Study and other relevant ones.

*Code and data are available at: LINK.

1.4 Recommended Uses:

Whom do you prepare the dataset for? Researching daily voting patterns, scientific studies on the nature of politics, knowing how demographics factor into election results, and creating models for predicting voter behavior starting from today.

2 Model Details - Manual

2.1 Model Details:

What is the function of a model? Forecasts voter attendance and choices using the demographic and historical information.

What type is the model? Seeking a supervised machine learning model, possibly a classification or regression model depending on the type of prediction task.

2.2 Intended Use:

Does the model cater the need for a specific gender or is it a gender neutral option? Politicians, campaign strategists, and social researchers in the political field.

What is the intended application? To forecast potential turnout and to contribute to understanding the factors of electorate casting their ballots for specific candidates.

2.3 Metrics:

How is that model going to be validated? Accuracy, Precision, Recall, F1 Score, and ROC-AUC curve for classification tasks.

2.4 Training Data:

What are model's training data? The US Cooperative Election Study Voter File of 2020.

2.5 Ethical Considerations:

What are the ethical concerns that may arise from this? Privacy issues, artificial intelligence bias in predictions, and the risk of those predictions misused for voter suppression.

2.6 Caveats and Recommendations:

What is the scope and accuracy of the model and data? The model's forecasting is determined by the historical data and does not take into consideration the current social movements.

What do we recommend this model to be used for? However, it is supposed to be the part of the research and analysis projects, not the source for predictions.

3 Ethical Aspects

3.1 Privacy and Anonymity:

Every data set of the individual voter may contain dangers of personal information disclosure. Anonymizing data is imperative because personal data shouldn't be claimed by anyone. Measures contain deletion of personal identifiers such as names, birthdays from the data or their replacement with hashes. In the event of disclosure. In addition, access controls should be implemented.

3.2 Bias and Fairness:

Voter and survey databases can simply hold societal biases, and the model can be used to discriminate. Consideration has to be taken to evaluate the model for equality among different demographic groups and incorporate methods that reduce any future effect of bias.

3.3 Misuse of Predictions:

Predictive models will be used, for instance, to aim at or get rid of some populations in the campaigns of voting, consequently leading to voter manipulation, or suppression among others. It is vital to set forth conspicuous directions and make a just standard for the use of model inferences to avoid unwanted applications.

4 The Dataset, Prediction and Model Tests.

4.1 Dataset:

Check for data completeness (e.g., invalid inputs).

Provide for the test of representativeness (e.g., the respondents come from a sample distribution that matches the population).

Test for veracity (like comparison with other reliable data).

4.2 Model test:

The cross-validation model to test for model robustness.

A/B testing in comparing model performance with baseline models.

Inclusion of bias and fairness testing tools to correct any inequitable predictions.

4.3 Predictions:

In a real-life situation, check (real-world validation test) the model by comparing its projections with the actual election outcomes.

Identifying the range of sensitivities of input to the projections of outputs.

Regular checking of the drifting in the data and the predictions delivered during the time to avoid any deviation.

Such a matrix setting would offer a platform that supports a prudent approach in voter behaviour data with a model driven by data.