2020 US voter file record*

Model card, ethical aspects, and tests for the model and dataset

Hyuk Jang

April 2, 2024

1 Model Card

Model Details

Person or Organization Developing Model: Hyuk Jang Model Date: April 2, 2024 Model Version: Up-to-date Model Type: Statistical model Training Algorithms and Parameters: Trained on the 2020 US Cooperative Election Study, post-stratified on an individual basis using a US voter file record from a private company. Features include demographic information, voting history, and other relevant factors. Paper or Resource for More Information: datasheet Citation Details: references.bib License: R Core Team (2023) Questions or Comments: NA

Intended Use

Primary Intended Uses: Predictive modeling of voter behavior, political analysis, demographic research. Primary Intended Users: Researchers, political analysts, social scientists. Out-of-Scope Use Cases: Individual voter targeting for commercial or political purposes without consent, discriminatory or unethical applications.

Factors

Relevant Factors: Demographic factors such as age, gender, race, voting history, socioeconomic status. Evaluation Factors: Accuracy, fairness.

Metrics

Model Performance Measures: Accuracy, precision, recall, F1-score. Decision Thresholds: NA Variation Approaches: Cross-validation

Evaluation Data

*Code and data are available at: https://github.com/anggimude/2020-US-Cooperative-Election-Study.

Datasets: 2020 US Cooperative Election Study, US voter file record. Motivation: Assess model performance and generalization. Preprocessing: Cleaning, feature engineering, post-stratification.

Training Data

NA

Quantitative Analyses

Unitary Results: NA Intersectional Results: NA

Ethical Considerations

Protection of voter privacy. Avoidance of discriminatory practices. Transparency in model development and usage.

Caveats and Recommendations

Exercise caution in interpreting results, as they may be influenced by biases inherent in voter file records and survey data. Regularly review and update model to account for changes in voter demographics and behaviors.

2 Ethical Aspects

Protection of Privacy: One important ethical issue with the features used in the model is safeguarding individual privacy. The dataset likely contains sensitive information about people, such as their demographics and voting history. While this data is crucial for understanding voter behavior, it also raises concerns about privacy breaches or misuse. To address this, strong privacy protection measures like randomization and encryption should be in place. Clear policies should also regulate who can access the data and ensure it's only used for legitimate research with proper consent.

Fairness and Bias: Another ethical concern is the potential for bias in the features used to train the model. Factors like race, gender, and socioeconomic status could unintentionally introduce bias, leading to unfair outcomes. For instance, if the model favors certain groups or perpetuates existing inequalities, it could worsen disparities in political representation or resource access. To tackle this, it's vital to thoroughly assess and mitigate any biases in the dataset or model. This might involve adjusting algorithms or carefully choosing features to ensure fairness for all individuals.

Transparency and Accountability: Ethical considerations also revolve around the transparency and accountability of the model's features. Stakeholders, including researchers and the public, should have access to information about the features and how they were chosen. Transparency builds trust and allows stakeholders to evaluate the model's decision-making process. Additionally, mechanisms for accountability, such as oversight committees and regular audits,

should be in place to address concerns or grievances related to the model's features or outcomes. Overall, transparency and accountability are crucial for maintaining ethical standards and ensuring responsible model use.

3 Testing

Dataset Testing:

Data Quality Assessment: Thoroughly examine the dataset to find and fix any oddities, errors, or missing information. This involves using statistical methods, exploring the data, and studying its characteristics.

Privacy and Security Review: Look into the dataset to spot any privacy or security risks. Put in place measures like hiding identities, coding data, and controlling who can access it to keep sensitive details safe and follow privacy rules.

Validation against External Sources: Check the dataset's accuracy by comparing it with other trusted sources or datasets. For example, cross-checking voter details with official records or census data.

Model Testing:

Model Validation: Make sure the model works well by testing it using different techniques like dividing the data into parts for training and testing, or repeating the training process with random samples. This checks how well the model can predict new data.

Bias and Fairness Analysis: Examine if the model gives fair predictions across different groups of people. Use tools to measure fairness and spot and fix any unfairness in the predictions.

Prediction Testing:

Evaluation on Test Data: Test how well the model predicts new data by checking it against a separate set of data not used during training. Measure how accurate the predictions are using various methods.

Comparison with Baselines: Compare the model's predictions with simpler models or common practices to see if it's doing better or if there's room for improvement.

Ethical and Social Impact Assessment: Think about the ethical and social consequences of the model's predictions. Consider if it might treat people unfairly or have unintended effects, and gather feedback from experts and those affected to address any concerns.

References

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.