Datasheet for a US Voter File Dataset*

STA302 Tutorial 12

Talia Fabregas

April 3, 2024

This data sheet discusses the motivation, composition, collection process, recommended uses, and more of a U.S. voter file record from a private company that has been used to train a model to predict the 2024 U.S. election.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The hypothetical dataset was created to enable analysis of the 2024 U.S.presidential election. While working on Term Paper 3, and now my final paper, I have not been able to find a publicly available survey data set that contains a large enough sample size and explicitly asks survey respondents about their preferred 2024 U.S. presidential candidate.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - I worked with the Eras Company to create this imaginary data set.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - Taylor Swift funded the creation of the data set. The grant name is the "Taylor Swift for President 2024 grant" and the grant number is 13.

4. *Any other comments?*

    - I have had to use data sets with ~1000 observations and no question about preferred 2024 U.S. presidential candidate for my previous research papers. This has been extremely challenging and I have had to make significant trade offs. I had to write code to infer preferred 2024 presidential candidate based on 2020 vote and

*Code and data are available at: https://github.com/taliafabs/tutorial12.git

party affiliation, which in most cases is obvious. However, when a respondent is an independent and did not vote for Trump or Biden in 2020, my 1000 observation data sets could not handle this because they were too small to fit a softmax regression model, so the only options were Trump and Biden.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - TBD
2. *How many instances are there in total (of each type, if appropriate)?*
    - TBD
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - This data set contains
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - TBD
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - There is a label or target associated with each instance (voter) in the voter file data set. It their preferred 2024 presidential candidate.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - TBD
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - Relationships between individual instances are not made explicit

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - There are no recommended data splits in particular

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - No

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - The data set is self-contained. it does not rely on external resources such as links, websites, tweets, or other data sets in any way, shape, or form.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - This is a voter file data set, so it does contain demographic and political preference information about each instance. For ethical and privacy reasons, we require all users to agree to the terms and conditions. This includes not sharing the data set publicly.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - This voter file dataset identifies the age, gender, race, and highest level of education of each voter.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

- It would be extremely difficult to identify individuals directly or indirectly from this data set. The voter files do include information about the state the respondent lives in, race, gender, age, and political preferences, but this information, even when combined with other data from external sources, is extremely difficult to connect to an individual.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - Yes. Race and political preferences are included in the voter file data set.

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - This voter file record was obtained from a private company. The data associated with each instance is self-reported.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Surveys and exit polls were used by the private company to create this hypothetical U.S. voter file record.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The private company used probabilistic sampling.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

- A private company was involved in the data collection process. Students at the Taylor Swift school interning for the company at the time were also inolved. They were paid $19.89 for each hour of work.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - The data was collected between 2020 and 2023.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No ethical review processess have been conducted as of now.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - I obtained the data via a third-party private company that has access to US voter file records.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - According to the private company, the individuals were notified about data collection via email.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Yes, the private company notified individuals via email.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - No, they were not.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- No

12. *Any other comments?*
    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - TBD
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - TBD
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R programming language was used to preprocess, clean, and label the data available. R can be downloaded via R project.
4. *Any other comments?*
    - TBD

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Yes. The data set was used to train a model on the 2020 US Cooperative Election Study and post-stratify it on an individual basis.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No
3. *What (other) tasks could the dataset be used for?*
    - This data set could be used to train logistic regression models to predict preferred 2024 presidential candidate based on race, gender, age, state, etc. Due to its size, it might be possible to use it to train a softmax regression (multi-class classification model) to predict whether individuals will vote for Biden, Trump, third-party, or abstain in the 2024 election.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - TBD

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - TBD

6. *Any other comments?*

    - TBD

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, the data set is available upon request for research purposes.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The data set will be distributed via GitHub. It does have a DOI.

3. *When will the dataset be distributed?*

    - After April 2024

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - NO

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - As of April 3, 2024, no.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No

7. *Any other comments?*

    - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The private company that owns this voteer file record will be supporting/hosting/maintaining the data set.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The owner can be contacted via email: imaginaryowner@privatecompany.com

3. *Is there an erratum? If so, please provide a link or other access point.*

    - No

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

    - This data set will be updated to add new instances once a month until October 2024. Updates will be communicated to dataset consumers via the mailing list that they join when they subscribe to the data set.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

    - Individuals were notified that the data would be retained for an unlimited amount of time

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Older versions of the data set will remain available via the private company's website with the date that they were updated.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - Those interested in contributing to the data set can contat its owner via email.

8. *Any other comments?*

    - No