# Datasheet

Alaina Hu

April 3, 2024

## 1 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 dataset was created to forecast election voting behavior and understand voting behavior of individuals. The task is to predict election results based on surveyed data. Election predictions in the past have often been criticized, so our analysis works to improve forecasts.

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

The Data Analytics team at Hu Private Company.

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.

No direct funding was received for this project.

4. Any other comments?

No.

## 2 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.

Each row of the main dataset is an individual of Hu Private Company. The row contains variables such as age group, gender, education level, race, political party affiliation, and candidate voted for.

2. How many instances are there in total (of each type, if appropriate)?

400 rows with 6 columns of variables.

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).

The dataset is a sample. It is a sample of employees at a private company. The sample is representative of the company population, but we are using this sample to cover the population of the country. In order to do this, we will use MRP.

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.

Each instance consists of information such as age, race, education, or political party, so these variables are all positive integers.

5. Is there a label or target associated with each instance? If so, please provide a description.

No, there is no unique key.

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.

All relevant information for this study is included in the data. No missing information for individual instances.

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.

Yes, each row represents the unique information of that individual.

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.

No.

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.

Self-contained.

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.

No, all data were gathered from individuals of the company with consent.

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.

Yes, each individual belongs to an age group, racial group, gender, political party, etc.

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.

No, not possible to identify individuals.

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, there is data on race and political opinions.

16. Any other comments?

No.

## 3 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.

The data was reported by subjects through survey responses. The data was not validated officially and self-reported by the subjects.

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?

Loading into R. All data was provided in the form of a csv file.

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 dataset contains randomly sampled data from the company. Every employee at the contact had equal probability of being selected to be surveyed.

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)?

Company employees. Company employees that agreed to be surveyed received a $5 Tim Hortons gift card.

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.

Data was collected over the period of one month. The survey was available for one month, and employees could choose whether or not to participate.

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.

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

Data was collected from the individuals in question directly.

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.

Yes. The data collection was through a survey that participants consented to participating in.

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. Participants knew that their survey reponses were to be collected and used for analysis.

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).

Yes, if participants changed their mind about their survey results being used, they could contact the data researches at any time before a given date to withdraw their information.

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.

## 4 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.

Yes cleaning of the data was done.

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.

Yes, the raw data sets are in the inputs/data folder, and scripts for the data cleaning are available in the scripts folder.

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 was used.

4. Any other comments?

No

## 5 Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

The exact analysis dataset has not been used for any tasks because it collected for the sole purpose of this study: election forecasting

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?

Other tasks related to voting behavior/elections.

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?

Legel risks. Future usage of the data would need to be reviewed and approved by the participants of the study. They were only informed of the usage indicated in this forecasting study.

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

No.

6. Any other comments?

No.

## 6 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.

No, the dataset will not be pubically distributed for purposes outside of the entity unless contacted and approved.

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

The dataset is private, but can be distributed with approval by the company. In that case, it will be distributed with an API.

3. When will the dataset be distributed?

Only with approval.

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.

Yes, the dataset compiled by Hu Private Company will be distributed under a specific intellectual property (IP) license, coupled with a comprehensive Terms of Use (ToU) agreement designed to protect both the dataset's integrity and the proprietary methodologies employed during its compilation.

License Type: Hu Private Company Open Dataset License (ZDODL v1.2)

License Summary: The ZDODL v1.2 is a custom license crafted to facilitate the wide dissemination of data while protecting Hu Private Company's proprietary interests. Under ZDODL v1.2, users are granted the following rights:

Use: Permission to use the dataset for both academic and commercial research only with approval. Modify: No rights to modify the dataset for personal research purposes. Distribute: Rights to distribute original or derivative works based on the dataset, provided that such distribution is non-commercial and includes clear attribution to Hu Private Company. Terms of Use (ToU): The dataset's ToU outlines the ethical and legal considerations associated with its use, including but not limited to:

Attribution: Required citation of Hu Private Company as the source of the dataset in all publications or presentations. No Warranty: The dataset is provided "as is," without warranty of any kind. Prohibited Uses: The dataset may not be used for unlawful activities, nor to support machine learning models in contexts that could result in harm to individuals or groups. Fees: Access to the dataset under ZDODL v1.2 is provided free of charge for academic and non-commercial uses. For commercial applications, a licensing fee is applicable, details of which can be obtained via our licensing portal.

Access Point: For more information on the ZDODL v1.2 license, the full ToU, and information regarding licensing fees, please visit our dedicated portal.

Contact Information: For further inquiries or to request commercial licensing terms, please contact our Licensing Department.

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.

Yes, certain datasets incorporated into the Hu Private Company compilation have been sourced under specific agreements with third-party entities, which impose additional IP-based and usage restrictions beyond those outlined in our primary Hu Private Company Open Dataset License (ZDODL v1.2).

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.

Due to the sensitive nature of some of the data compiled by Hu Private Company, certain datasets and individual instances are subject to export controls and regulatory restrictions under both national and international law. These controls are implemented to prevent the unauthorized dissemination of sensitive technology, data, and information that could compromise national security, privacy, and proprietary interests.

7. Any other comments?

No.

## 7 Maintenance

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

Alaina Hu

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

alaina.hu@mail.utoronto.ca

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)?

No, the dataset will not be updated.

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.

Yes, the applicants were told that their data would be retained for 10 years and then deleted. They consented to this time frame. The data is secured and will only be available with approval and licensing.

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 of the dataset do not exist right now in Hu Private Company. This is new data solely for the purpose of the forecasting.

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.

No, the dataset will not be extended. For the efficiency of the study, participants had one month to contribute their data for this study. After the month, the survey closed and the dataset no longer accepted contributions.

8. Any other comments?

No