Stor 390 final project

Eric Rash

2024-04-26

Leveraging Digital Trace Data for Electoral Predictions: A Critical Analysis

In a time characterized by the influence of developing technologies, the potential of digital trace data in predicting electoral outcomes has gotten some attention. Digital footprints left behind by individuals navigating the World Wide Web offer a promising avenue for understanding voter behavior and preferences. The allure of real-time insights into voter sentiments and trends is undeniable. Supporters believe that leveraging digital trace data could revolutionize electoral forecasting, providing nuanced understandings of voter behavior while democratizing access to information. Conversely, skeptics caution against the use of digital trace data, citing concerns regarding privacy, and algorithmic biases. With this context, this paper will be reviewing a study that attempts to do exactly what has been proposed: predict voters in an upcoming election by using digital trace data. This paper is structured into four distinct sections. The first section, Analysis of Methods, will delve into the methods used throughout the paper while offering novel insights via simulated statistics through similar methods. By understanding the strengths and limitations of these methods, I argue against the usage of digital trace data for these purposes. The second section, Analysis of Normative Considerations, will evaluate the ethical dimensions of utilizing digital trace data in predicting electoral outcomes. Drawing upon deontological principles discussed in class, I will assess the implications of such practices in terms of the rights of persons, and the universality of these practices. By engaging with both methodological nuances and normative considerations, I hope to contribute to informed discussions surrounding the intersection of technology, politics, and philosophy. This analysis underscores the imperative of adopting a responsible and ethically grounded approach to the utilization of digital trace data in electoral predictions and the importance of adopting ethical frameworks for the implementation of new statistical methods in any context.

Analysis of Methods

```
The authors attempt to predict how people will vote in an upcoming German election using digital trace d
```

Undecided - Did they respond that they were undecided in surveys? *Voted* - Did they respond that they voted in surveys? *AFD* - Did they respond that they voted for this party? *Greens* - Did they respond that they voted for this party? They chose the most polarized parties because they believe having more polarization between the two parties will make it easier to classify an individual based on their digital trace data. They have three different block predictors with varying predictive abilities. The first includes information about the device and when it is used. The second includes the duration and frequency with which respondents access the 50 most used news domains. The third block of predictors features websites/apps that were used more than 80 times for over a minute. After combining everything there is a total of 11999 predictors used by their algorithm from the digital trace data. When creating their model, the researchers decided on using XGBoost because it can filter through predictors to select only the best variables in the model-building process. These create various decision trees, then use the issues from the previous tree to create a new

tree that should be an improved model. This model is extremely good at classification and ranking. Better put, the XGBoost is capable of trimming down a model so that it is small enough to be interpretable and retains strong predictive powers. Allowing for users to reach the simplest model that will still produce accurate predictions. The researchers created, "a total of 35 XGBoost models, 6 for each of the political outcome variables (undecided, voted, AfD, and Greens) and 1 for each of the sociodemographic outcome variables (age, gender, net income, marital status, federal state, childless, number of children in household, and employment status)." (Bach, L. R., et al.2019). They add that these groups allow them to gain insights at different levels. Group 1 just has demographic data, group 2 just has digital trace data, groups 3-5 have different combinations of each, and group 6 contains everything. The researchers find that their models are unable to perform at the level of typical socio-demographic models made currently. They add that digital trace data struggles greatly at predicting how undecided voters will vote, but it does do a better job with more populist parties. They find that they are unable to predict how someone will vote by using digital trace data. They add that they can consistently predict the age and gender of respondents using their digital trace data. Additionally, they believe that if they had more resources to devote, like a major company like Google, they may have better luck at predicting these measures.

Analysis of Normative Concern

As mentioned in the introduction, I will analyze this paper from a deontological perspective. Before I
In the context of the paper I am looking at, a deontologist would likely be against the usage of digita
Using digital trace data violates individual autonomy. In the case of this study, their data was acquir
Finally, there are transparency concerns with the usage of this data, specifically as it applies to mac
The universality concern can manifest itself in a couple of ways in the context of digital trace data.

Conclusion

Digital trace data is already being used to fundamentally shape every user's online experience. Algorit