Model Card*

STA302 Tutorial 12

Talia Fabregas

April 3, 2024

This model card outlines the details, intended use, factors, metrics, and intended data of a logistic regression model trained to forecast the 2024 U.S. presidential election.

Model Details

- This model was developed by Talia Fabregas with the guidance of Prof. Rohan Alexander and the material presented in STA302
- Model date: April 2, 2024.
- Model version: Talia's Version
- *Model type* : This is logistic regression model built using the `stan_glm` function and default priors of the `rstanarm` package.
    - However, if we want to consider the multi-class case (where not voting and voting third-party are options), a softmax regression model can be built.

Intended Use

- The primary intended use of this model is for it to be applied to post-stratification data to forecast the 2024 U.S. presidential election.
- The primary intended users are statisticians, researchers, and students interested in this subject matter.
- Out-of-scope uses: Not for use on training data sets with fewer than 10,000 respondents or census data sets from before 2022.

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

Factors

Metrics

Model performance metrics -

Evaluation Data

- The 2020 U.S. Cooperative Election Study was used to post-stratify
    - It was publicly released after the 2020 U.S. election (the CES is always released after the presidential election), so we must use the previous one to help forecast the 2024 election.
    - According to Telling Stories with Data, this is the best statistical option, despite the fact that it is only available after an election.
- Motivation: According to Telling Stories with Data, it is the best statistical option. This is also the one provided in the context of STA302 tutorial 12. It provides robust census data about voters in the United States.
- Preprocessing: This data set was pre-processed for evaluation by selecting relevant variables and re-naming those columns to match my training data.

Training Data

- The dataset used to train the model is a U.S. voter file from a private company was used to train our logistic regression model to predict 2024 vote choice based on state, age, race, gender, highest level of education, marital status, and whether the respondent lives in an urban or rural area.
    - The predictors used in this model nearly mirror the ones that I used in term paper 3.
- Motivation: This data set was chosen because it contains more than 10,000 voter files and a variable indicating preferred 2024 presidential candidate.
    - Unlike most existing survey data sets, it does not require us to infer preferred 2024 presidential candidate based on 2016 and 2020 vote choices and party affiliation.
- Preprocessing: this data was pre-processed for evaluation by selecting the columns corresponding to 2024 preferred candidate, state, age, race, gender, highest level of education, marital status, and urban or rural area.
    - The voter file contains much more information than that, so we subsetted it to only include the information for each individual in question required to train the model.

Quantitative Analysis

- Unitary results: In term paper 3, I had warnings about my training sample size being too small. I expect that issue to resolve now that the voter file contains 10,000 observations. I expect to see high training accuracy because the variables I selected are known to predict vote choice.
- Inter-sectional results: I expect to see the model perform well with respect to intersectional factors.

Ethical Considerations

- Data: The model uses sensitive data, including race and gender. Race and gender are protected status.
- Human life: This model is not intended to inform decisions about matters central to human life or flourishing, nor do we believe that it can be used in such a way.
- Mitigations: Risk mitigation strategies that can be used during model development include discussing the weaknesses and limitations and ensuring the privacy of survey respondents.
- Risks and harms: risks and harms present in model usage include the creation of misperceptions about the upcoming U.S. presidential election. It is important to know that this is just a model, and it has weaknesses. This cannot be treated as a clear idea of what will happen in November.
- Use cases: Known model use cases that are especially fraught include the use of the results in a short social media post or as clickbait online without offering any additional context, such as how it was developed, trained, and what predictors were used.

Caveats and Recommendations

- Additional testing, analysis, and research is needed.
- As seen in Tutorial 10, there is no clear-cut way to choose predictors for the logistic regression model. It may be wise to consider slightly different logistic regression models and compare the results.