IMPACT OF SOCIAL CONNECTEDNESS ON POLITICAL PARTICIPATION

MPCS 53120: Applied Data Analysis

Emily Ding

June 2, 2022

Executive Summary:

This project uses county level social connectedness data from Facebook's Data for Good as well as demographic factors like race, education, income, and gender to predict voter participation in the 2020 election. The work begins with sourcing and engineering data, in particular, pivoting social connectedness data into distance buckets, normalizing into proportions, and joining datasets together by county. Then, models are generated using OLS linear regressions, Forward Step Feature Selection, Lasso Regression, General Linear Modelling, and XGBoost with Randomized Search parameter tuning. Models are compared by using root MSE scores to select for the most effective model. Finally, comparisons are made between models containing Social Connectedness Data as well as those without it to examine whether social connectedness has any predictive power towards vote participation.

Introduction:

Our social connections form an integral part of the human experience. From a political standpoint, political scientists have theorized extensively on the role social connections play in shaping political engagement. The people we interact with impact the resources we have access to, they serve as our connection point to institutions of power, and they directly expose us to ideas that sway our opinions on matters. From the late 1990s, political scientists began to express concern over increased social isolation and its detrimental impact on the health of our political system, theorizing that more socially isolated individuals are at a danger of not engaging. Subsequently, theorists cemented that the robustness and form of social networks have a direct influence over political engagement. These ideas made their way into models.

A major challenge in this area, however, is in not only identifying which predictors are meaningful, both in a theoretical and a measurement sense, but also in identifying an appropriate proxy for something as intangible as a social network. Recently, however, the rise of social media, with the digitization of relationships, offers a new solution. For my project, I would like to examine whether social media data, particularly the Facebook's social connectedness index data, can help predict political participation.

Literature

From my survey of the literature, this is a relatively new data set that has not been applied to predict political outcomes. Because of this, I mainly consulted the literature to source methodological suggestions.

First, I was curious what existed in the literature about the impact of social isolation on voter participation. Jack Lyons Reilly, in his paper in Research and Politics, found that "isolated individuals clearly participate in politics less than individuals who are well connected to those around them." In his analysis, he used the 1992 CNES-A in which social connectedness and political participation were generated by randomly selecting individuals to name five individuals that they discuss political networks with. He regressed these against metrics such as turnout, various issues, and voting stance to look for trends. Most relevantly, he found that highly socially connected individuals' turnout rates are 24 percentage points higher, while social connectedness did not influence how people voted. While my project takes a more geographical interpretation of "social isolation", this paper establishes the precedent for using social connectedness in political participation models.

Next, I looked for suggestions on how to consume the facebook SCI dataset which I wanted to use. The establishing paper on this was by Bailey, Cao, Kuchler, Stroebel, and Wong in the Journal of Economic Perspectives. It is in their work that I discovered the method of using Haversine distance formula to translate location pairs into distances and then categorize number of friendships into distance buckets of less than 50 miles, 50-100 miles, 100-200 miles, and 200+ miles. The team had used both probabilities and elasticities to qualify each location before running regression models. I referenced this methodology when constructing my models.

Lastly, from Timpone in Ties that Bind, we get a survey of the demographic predictors that have been theoretically and quantitatively shown to be useful in political engagement predictions. These included region, education, age, gender, race, income, marital status. Timpone aggregates these predictors that from what have been proposed and used throughout literature and evaluates their connection to his proxy of Social Connectedness. I reference these when constructing my model. The authors found that while education, income, and age were related to social connectedness, including social connectedness only weakly changes the impact these variables had on political participation. These made sense to be used as a predictor set.

Data Engineering:

My first big challenge was in engineering the data set that I needed to construct my models. The most interesting and important data set that I wanted to incorporate came from Meta (Facebook's) Data for Good initiative. A team of data scientists at Meta partnered with the Journal of Economic Perspectives to consolidate data into what they coined the "Social Connectedness Index." This resulting data set presented pairs of counties/zip codes/countries and the number of Facebook friendships that exist between users residing in two locations. The data points included in the dataset are cleaned for user privacy.

For my project, I used the county data set from 2020. I made this decision for a couple reasons. First, since I am exploring political activity, many precincts and political engagement initiatives are organized at the county level. This is granular enough to allow for variation but aggregated enough to make it feasible to work with. Second, counties are organized by the US Census into FIPS codes, a five digit code encoding state and county identification. This reference is consistently used by other data collectors, mainly the US census as well as election and poll level data sets. By having FIPS codes at hand, I was able to join multiple data sets from different,while assuring that location boundaries were consistent.

Before I could interpret the county level social connectedness data, I had to reorganize it so that county, not county pairs, were the indexes. In order to do so, I needed a way to characterize each county's social connectedness makeup. I borrowed methodology from the above discussed Journal of Economic Perspectives article and ratioed the social connections in each county into their respective distance buckets. The team had applied the Haversine distance formula between counties and then categorized all the friendships into buckets of "less than 50 miles", "between 50 – 100 miles", "100 – 200 miles" and "200 and up miles." These buckets

made intuitive sense as well, since, as an example, 50 miles is roughly the distance from Chicago to Naperville, while 200 miles would be Chicago to Ann Arbor. While I used these buckets to continue off the literature, I'd be interested in further studies that categorized differently, perhaps by using different distance boundaries, such as 300-400 miles, 400-500 miles, etc or more qualitative boundaries such as state or region (Midwest, Pacific West, New England). Instead of calculating the distances between counties manually, which I found difficult to do accurately without determining a specific coordinate of reference for each county, I decided to expediate this process by referencing a dataset from the National Bureau of Economic Research which provided the county pairings and their distances calculated using the same Haversine formula. I joined the social connectedness dataset with the distance dataset by FIPs code to find the distances between each county pair.

There were a couple points of loss from this approach, which I would like to mention. First, I had to drop a set of counties where the distance data was not available for, predominantly those that were outside of the landlocked states. In particular, these were the territories of American Samoa, Northern Mariana Islands, Guam, and Virgin Islands. This is not much of a problem as I would have dropped them later, since the territories do not contribute electoral college votes. Secondly, the distance data set is from the 2010 census. I had a lot of difficulty finding a more up to date data set, and so ultimately, while not ideal, I did some manual reassurance that county boundaries have not changed a whole lot since 2020. With distances now identified, I categorized each county pair into one of our distance buckets of "less than 50 miles", "between 50 – 100 miles", "100 – 200 miles" and "200 and up miles."

In order to flatten the data so that it is indexed by county (and not county pairs), I used the pivot table functionality in pandas to aggregate all the number of connections into each

county by distance bucket. For a given county and distance bucket, this summed all the connections. To make counties comparable, I opted to calculate a proportion per each bucket. The final form of our SCI data is such: for a given county, we now know the proportion of connections less than 50 miles away, 50-100 miles away, 100-200 miles away, and 200+ miles away. It is worth noting that these buckets add up to 1, which we will need to handle later to avoid collinearity. From an interpretation standpoint, these ratios described the makeup of a counties isolation level – counties with a high portion of connections in the "<50%" bucket were less connected outside their immediate geographic area. This is a more geographical definition of isolation compared to the original literature of individuals who did not have as many social connections. Regardless, it is an interesting quality to study.

The next set of data that I needed to bring in was the dependent variable. There were a couple avenues I could take to examine political participation – from voting, such as whether or how people voted, to community engagement, number of referendums, frequency of candidate lawn displays or bumper stickers (per what the American Community Survey asks about). From the literature, political scientists were particularly invested in engagement trends. For this reason, I focused most of my analysis on whether people voted. The data set I used to measure this came from the 2020 presidential election. This is a dataset I found on github that reported each county's FIPS score, the total number votes, as well as the number of votes towards the democrat vs. republican candidate. Since raw number of votes needed to be normalized against the county's voting population, I found a dataset giving each county's estimated CVAP – or citizen voting age population. Dividing number of votes by estimated CVAP gave what I called percent participation in the election for a county. Note, this is different from turnout, as turnout is conventionally based off of number of registered voters. I decided to go with proportion of

voting age population as people who would be truly disengaged may not even register to vote. Since I wanted to use linear regressions later, I plotted the data in a histogram to check for normality. The data indeed made a rough bell curve. A concern that I had about the data is that for a couple (~6) counties, the participation was a bit over 1. This is not possible in the real world but exists because CVAP is an estimated number. I considered dropping them but was worried it would skew my data. Since they were not too much over 1, I decided to keep them in.

Lastly, I sourced multiple sets of demographic data to use as additional predictors. These I mainly found from the US department of agriculture and the US Census. I picked data representing race, education, gender, and income levels. These were the main predictors from other political participation models from literature. After cleaning the data and converting values to percentages, I was able to merge these into one large dataset using county FIPs code.

Modeling:

The most straightforward starting point in modeling is with OLS regressions. But, first and potentially most concerning, we must deal with the fact that the outcome variable (participation) is a proportion. This means that the values should be constrained to between 0 and 1, which is not a bound that linear regression adheres to. Additionally, because the range of values of percentages are quite small, it was imperative to add an intercept to any regression model that I worked with. Lastly, as mentioned above, the distance buckets always add up to 1. In order to not introduce collinearity, I dropped at least one bucket from each categorical set when running the models. It didn't matter which buckets were dropped, but since the more extreme buckets were more interesting when interpreting results, I usually chose one of the middle ones.

The first thing I was interested in modeling is whether SCI had any predictive power towards political participation. Just from a scatter plot of the different buckets against participation, we can visualize a weak relationship, but it was unclear if it were linear.

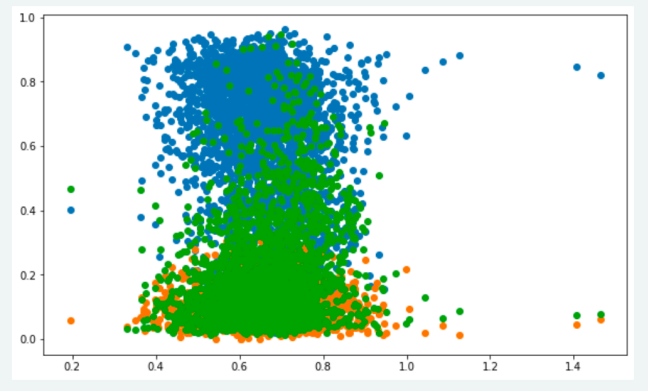

Figure 1 Distance Buckets vs Participation

As a sanity check, I checked the VIF values for multicollinearity. The values were fairly small (<5), so I felt it reasonable to continue running an OLS regression of the distance buckets against participation. The adjusted R squared value was pretty dismal (0.045) but the coefficients were significant at a 0.05 level, except for "%200+" miles bucket. I reran the OLS model to see if just two distance buckets were enough. In fact, for the same adjusted R squared, two distance buckets was. The F statistic was also high (391.7), leading me to believe that while SCI was not a powerful enough predictor of vote participation on its own, it does have predictive power in a model.

To that end, my next regressions involved introducing demographic explanatory variables, such as race, education, gender, and income. Off the bat, I started with an OLS regression using all the variables I had – barring the ones I dropped from race, education, gender, and distance so that they did not add to 1. The adjusted R squared value and the root MSE was much improved from the SCI only regression, and all the predictors looked significant. However, with such a large set of predictors – which was expanded because race, education, and gender were categorical variables— I was worried about overfitting the model due to strong correlation between variables. I knew some of these, like race and education, have historical data that shows strong correlation between them. In fact, checking the VIF definitely shows so. Percent of White Americans (%WA), for example, had a VIF of over 600!

Thus, the next challenge then is how to select for the right set of predictors. I first used a forward stepwise function and picked the optimal model by adjusted R squared value. This seemed like a viable way forward as it added predictors one by one to see which additional predictor would improve the model the most. This turned out to be the model with 7 predictors – gender (% female), median income, a couple categories from education, Asian American and

American Indian/Alaskan, as well as <50% distance. What is most interesting is that one of the distance categories was included in the model. I went through (not shown in code for conciseness) and examined the 1, 2, 3, 4, 5 predictor models that the stepwise function identified. It was the 5th predictor added after gender, median income, and Bachler/Some college education buckets. One drawback that I did not get a chance to explore is that the stepwise function was performed without the intercepts. If I wanted to go further down this route, I would have vetted the stepwise function with intercepts, if possible.

As an alternative method, I used lasso regression to pick predictors, to encourage a more simple and sparse model with less predictors. The resulting linear regression model did not include the distance buckets. This is a deviation from our results in terms of the predictive ability of SCI – the improvement to the model introduced by the distance predictors might not be worth the additional complexity of the model.

Moving forward, we now consider models not under the linear regression family. I saw suggested on multiple online threads that General Linear Models with a logit link and binomial family was recommended for modeling continuous proportions. This also worked well as from a plot, it did not seem like the errors were exactly normal, which is a case that GLM binomial handles better than linear regressions. The logistic link helps keep the data bounded between 0 and 1. As a caveat, I originally considered using a Logistic Regression; however, my outcome variables were not binary. I could not use the model to make predictions. The GLM allowed me to constrain the outcome to 0 through 1, with the outcome a continuous value within that range. I was curious if I had to do feature selection for GLM models, so I attempted to apply stepwise selection to GLM. However, this was pretty futile – the RMSE improved as each predictor was added.

The model that I had the most success with, ultimately, was using XGBoost. XGBoost was a highly praised and powerful tree boosting machine learning package. It has built in functionality to smooth weights and avoid over-fitting. While I still have the same problem of bounded continuous proportions as my dependent variable, I also have to worry less about proportions as XGBoost is a tree based model, so the predictions it generates will generally fall within the trained data range. I first tried using an alternate booster type (gblinear) but the default of gbtree performed better. Upon analyzing for feature importance, the "under 50 miles" SCI bucket was by far the most important explanatory variable. This definitely made the case that SCI was an important and useful predictor of political participation. To make the XGBoost model as effective as possible, I took a recommendation to use the RandomizedSearchCV library to hypertune my parameters to find the optimal set. This gave me my best RMSE yet of 0.05575.

Last, for each model, I reran with and without SCI predictors so I could compare whether introducing SCI predictors improved the predictive power of my models.

Results:

To compare models, I referenced each model's root MSE in predicting a 10% test data set. Note that the units of this measure is in proportions, so the magnitude of the number is very small. The tuned XGBoost Model performed the best.

```
{'OLS w/ 3 SCI buckets': 0.09493309653983951,
 'OLS with 2 distance buckets': 0.09490857101592567,
 'OLS with all demog predictors': 0.06854140161179204,
 'Forward Stepwise w/ 8 predictors': 0.06996915351642546,
 'Forward Stepwise w/ 7 predictors without SCI': 0.07080617020376986,
 'Lasso Regression': 0.07080617020376986,
 'GLM': 0.06902116966029863,
 'GLM without SCI': 0.07056290805588866,
 'tuned XGBoost': 0.055750325457463855,
 'tuned XGBoost without SCI': 0.05645178476638168}
```

Figure 2: RMSE Results Summary of All Models

In addition, my conclusion is that while weak, Social Connectedness does help predict political participation. In every model, the addition of SCI predictors improves the model. In particular, our optimal model, the tuned XGBoost model, identified "<50%", "%200", and "%100-200" buckets as the top three most important predictors.

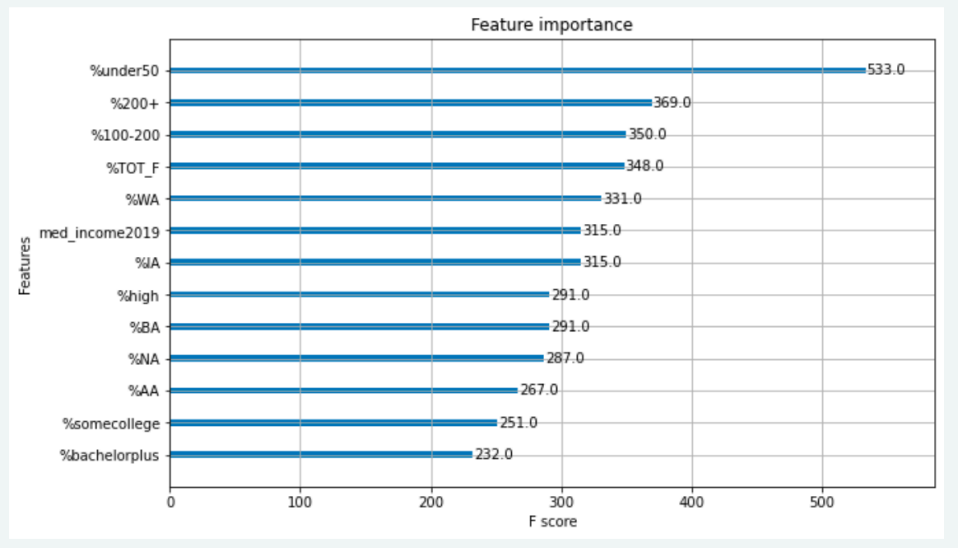

Figure 3: Feature Importance for XGBoost Model with SCI

Effort:

For this project, I divided my effort into quarters, each of which took 15-20 hours. The first quarter was thinking of ideas, idea refinement, reading about the topic, reading related literature, and writing the proposal and mid quarter presentation. I spend a second quarter of my effort on data sourcing, cleaning, and organization. This was the part that took longer than originally anticipated because I had to bring in and join data from many sources. Often I was able to find the data but not in the form or organization I wanted, and I wasted some time trying to scrape and engineer data points that I end up scrapping. Ultimately, limiting my data sources to ones organized by FIPs helped me a lot with consistent matchup of data. For this project, I learned about pivoting and merging data in order to organize my data in a way that was usable. The third quarter of my effort was on model creation, optimization, and researching. I started with a foundation of what we learned in class, namely OLS models, feature selection, and Lasso regression. Then, I learned more about GLM binomial models as well as learned how to use XGBoost. XGBoost is a new library that I heard about through our project discussions with Dave. I also had to learn how to tune xgboost parameters. I dedicated some time learning about what people used to work with and handle data that is continuous but bounded – in my case, proportions. My last quarter of time is spent on paper writing, presentation preparation and recording, as well as citing my sources.

Future work:

While I was working on this project, there were a couple areas that I would have liked to look into further. These I broke down into 1) better data engineering 2) additional models 3) better optimization 4) additional theoretical explorations.

To start, there are points in my process where I would have liked to acquire better data. For one, I am using participation, but I would like to explore whether trends are more enhanced with voter turnout data. Since I could not find a centralized repository for turnout, I would have needed to scrape state/county voter registration data to get this information. Second, I felt that the facebook social connectedness data was such an interesting data set, there were a lot more insights that could have been gathered by constructing the data differently. This could have been by different distance breakdowns. Facebook also has data at the zip code level and the state level, which may lend itself to more interesting nuances. Particularly, I felt that county could be too broad at times. If you consider that the entirety of Chicago is one county, it is likely that trends at the neighborhood level would be lost. Third, I have some concerns about data quality that I would like to fix by finding better data. Through the course of the project, I've aimed to join data from very similar time periods (~2020); however, some of the data was a bit older (2010, etc) or was aggregated (2016-2020). This discounted the integrity of my models.

The next area of additional work to explore would be additional models. I was unfamiliar with working with proportional outcome data, so I know that I only scratched the surface in terms of handling this type of data. As an example, I began reading about Beta Regressions, which were another recommended way for handling this type of data. I did not find a library to run this type of regression but it was something that people have manually coded. With more time, I would love to learn about and implement this regression type. In addition, I have

discovered the power of gradient boosted models! XGBoost was a popular one, but there were more out there that I would be interested in learning about. These were very effective!

Third, I had a good discussion with Dan in office hours about parameter tuning libraries. He told me about libraries like Optuna, which although computationally expensive, can be implemented to tune parameters even more than I was able. This is pretty cutting edge and very interesting – in the future, I would love to explore this path further. I used random search to tune some of my parameters in XGBoost, but it would be much more effective for an automated library to iterate through all combinations of parameters and identify the optimal.

Lastly, there are additional theoretical follow-ups to explore. I considered this in my proposal, but did not get a chance to explore additional political metrics – such as how people voted or vote difference margin. It would be a natural follow-up to explore more than just one election or even look into trends over time. My data is snapshot, so it is from a single point in time, but I would be curious in seeing how changes in social network makeup over time influences our results. I also think there would be a lot of insights that could be garnered from comparing between election cycles (2016, 2012, etc) and election type (primary, presidential). Finally, social isolation and connection could be effected by more than just location. It would have been an even stronger analysis to examine how many friends people had. I believe this is something that can be engineered from social media data, however, as of now, this data has not been made available.

References

Bailey, Michael, et al. "Social Connectedness: Measurement, Determinants, and Effects." *Journal of Economic Perspectives*, no. 3, American Economic Association, Aug. 2018, pp. 259–80. *Crossref*, doi:10.1257/jep.32.3.259.

"Citizen Voting Age Population by Race and Ethnicity." *Census.Gov*, United States Census Bureau, 17 Mar. 2022, https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/cvap.html?fbclid=IwAR2zDh-uvqjVd7R7JuPUqFt5I1KTNsYapfXZmRR8wLVv8gy6yGWfFm0AX4Q.

"County Distance Database." *NBER*, National Bureau of Economic Research, https://www.nber.org/research/data/county-distance-database?fbclid=IwAR2cBBZEDyKjGeRdRYxcLvNQ-Adtimof-c90WbGo9NVXd4G9disZ_cFFGxI. Accessed 31 May 2022.

"County-Level Data Sets." *USDA ERS* , US Department of Agriculture, 2020, https://www.ers.usda.gov/data-products/county-level-data-sets/?fbclid=IwAR3GVM_rOE0vYDn0srU9uFl9aZd3FeEzrdFwLUgsfqq2dGkB0eGTpc5u1II.

Crawley, Michael J. *The R Book*. John Wiley & Sons Ltd., 2007.

"Facebook Data For Good Social Connectedness Index Methodology." *Data For Good Home*, Meta, Aug. 2020, https://dataforgood.facebook.com/dfg/docs/methodology-social-connectedness-index.

Krishna, Harish. "XGBoost: What It Is, and When to Use It." *KDnuggets*, https://www.facebook.com/kdnuggets, Dec. 2020, https://www.kdnuggets.com/2020/12/xgboost-what-when.html?fbclid=IwAR3ShFqEMiYGhOrmknyUumjgjWLKrUcyh0URU_lTun2Y6JYQmmGjKvQnD0Y.

MegMeg. "Interpreting Proportions That Sum to One as Independent Variables in Linear Regression." *Cross Validated*, Stack Exchange, Dec. 2015, https://stats.stackexchange.com/questions/183601/interpreting-proportions-that-sum-to-one-as-independent-variables-in-linear-regr?fbclid=IwAR29OcBIK38P78PC7py_TtwUj1bf7aTejepQTM_-KKudYYrPFFlkiZzLcFw.

NizagNizag78911. "Variance Inflation Factor in Python." *Stack Overflow*, Stack Overflow, Mar. 2017, https://stackoverflow.com/questions/42658379/variance-inflation-factor-in-python?fbclid=IwAR3wyet8m8r159ewcvos1tekFKe_56C7Ak-ZBP0WkvwbN3NBoGijMP-pvm8.

Reilly, Jack Lyons. "Social Connectedness and Political Behavior." *Research & Politics*, no. 3, SAGE Publications, July 2017, p. 205316801771917. *Crossref*, doi:10.1177/2053168017719173.

tonmcg. “US County Level Presidential Results.” *GitHub*, Aug. 2020, https://github.com/tonmcg/US_County_Level_Election_Results_08-20/blob/f9b5f335ad1c66a7eba681539db49eec0c22787b/2020_US_County_Level_Presidential_Results.csv?fbclid=IwAR1EIEh5wBGWYqPZ-I57eMmiP6B28SWf6JuMw7myTXf6VVUTtVuQwvLoeqI.

“XGBoost 1.6.1 Documentation.” *XGBoost Documentation*, dmlc, 2021, https://xgboost.readthedocs.io/en/stable/parameter.html?fbclid=IwAR2FVvpojJ7faZQ3_Iy0gzbWpT9i9lRLUamZAAs0iRD1nnvvy7ZmcFDEhHA.