Internet usage and Interest in Political topic

Justin Tien

SODA 496

Professor Michael Burnham

05.05.2023

Logo

Description automatically generated

**Introduction.**

Our research project aims to examine the relationship between internet usage, social media, and voter turnout in the 2016 and 2020 US presidential elections. Specifically, we want to investigate whether there is a significant difference in political engagement levels between the two elections, and whether social media played a role in any potential increase in engagement. Given the ubiquity of social media today, it is important to understand how it affects political participation and potentially influences election outcomes.

**Theory and Hypothesis.**

The modern democratic system relies on elections as a key mechanism for determining political leadership, and during election years, people tend to be more interested in political topics due to the direct impact on their daily lives. With the increasing use of social media, politics is being discussed more frequently on these platforms, leading to potential positive and negative outcomes. While many Americans engage in politics during election periods, it is unclear whether they are willing to share their political opinions on social media. Therefore, our study aims to investigate the relationship between social media use, political engagement, and voter turnout in the 2016 and 2020 US presidential elections.

Hypothesis: There will be a higher level of political exposure in the 2020 election compared to the 2016 election, due to increased social media use.

**Literature Review.**

In this study, we aim to explore the relationship between social media usage and voter turnout during the 2016 and 2020 elections. Social media has become increasingly popular in recent years, and it is now being used to discuss political topics more than ever before. While some studies have looked at the impact of social media on political participation, there is still much we do not understand about the relationship between the two.

Previous studies have shown a positive correlation between internet use and political participation, indicating that internet use can directly affect the amount of political exposure to users (Fuchs, Trottier, 2015).

Prior research has focused primarily on Facebook and the impact of social pressure on political participation (Haenschen, 2016). By examining social media activity in relation to primary election turnout rates, we hope to gain a better understanding of the impact of social media on political participation and whether it played a role in the difference in engagement levels between the 2016 and 2020 elections.

**Data Summary.**

To measure internet usage, two databases are utilized: "Internet usage penetration in the United States in November 2021, by state" from the Computer and Internet Use Supplement to the Current Population Survey (CPS) conducted by the US Census Bureau, and "Number and percentage of households with computer and internet access, by state: 2016" from the National Center for Education Statistics. Both databases provide data on internet usage by state during the primary election year, with 50 observations for each database. Voter turnout is used to measure political engagement and interest, and two databases are utilized: "Reported Voting and Registration for States: November 2020" and "Reported Voting and Registration of the Citizen Voting-Age Population, for States: November 2016." Both databases contain data on the registered population and number of people who voted. Other variable including poverty, winning margin and battle ground state data also include to exam model accuracy.

**Inferential Model.**

We are developing an inferential model for the 2016 and 2020 primary elections. Our analysis includes generating pairplots to identify the optimal set of features that explain the relationship between two variables or to form the most distinct clusters. We are then constructing linear regression models and OLS regression models to obtain the R-squared score, which quantifies the proportion of variation in the dependent variable that can be explained by the independent variable. Additionally, we are using the F-statistic and P-value to determine whether the means between two populations are significantly different.

Chart, histogram

Description automatically generated Chart, scatter chart

Description automatically generated

**(2016 inferential pairplot)**

Chart

Description automatically generated Chart

Description automatically generated

**(2020 inferential pairplot)**

**Predictive Model.**

In our predictive model, we utilized the train\_test\_split method from the sklearn.model\_selection library to construct the model with a majority class classifier, along with measures of mean and model score. Additionally, we performed data refinement by dropping different variables.

To identify the most influential features for predicting the target variable, we used the technique of feature importance. Specifically, we used logistic regression with single variables and compared mean and model scores to select the most relevant input features. Furthermore, we conducted a feature importance analysis using random forest models, building various models such as the full model, model without poverty, MR, and BG data, and model without internet data, among others.

**Result.**

**Inferential Models:**

We build the linear regression model by X = internet usage and other variables (poverty, winning margin, and battleground data), Y = level of state’s voter turnout (1 = higher which is greater than mean, 0 = lower which is less than mean value).

For the 2016 model, the intercept was estimated at -3.06, and the coefficients for each variable were as follows: households with computer access (0.003), internet access (0.042), poverty (-0.023), winning margin (0.0005), and battleground (0.12 for Y=1, N=0). The variables with the highest coefficients were battleground data and internet access, but their values were not significant enough.

For the 2020 model, the intercept was estimated at -1.05, and the coefficients for each variable were as follows: internet access (0.018), poverty (-0.097), winning margin (0.0091), and battleground (0.654). The variables with the highest coefficients were also battleground data and internet access. However, BG variable value may not be accurate, because both Y and BG variables are both binary values.

For the R square value, the 2016 model has R square = 0.231, which means 23.1% of the variability in the outcome data can be explained by the model. 0.23 would show a low correlation.

2020 model R square = 0.443, which means 44.3% of the variability in the outcome data can be explained by the model, which is higher than the 2016 model, however, it still does not have a high correlation.

Moving on to the F-statistic value, the 2020 model has a higher value (9.153) compared to 2016 (2.711), which can be interpreted that the 2020 model is better than 2016.

Internet access has the highest correlation among all the variables except battleground state data, however, it may be inaccurate because the battleground state and the label of voter turnout both are binary values.

To sum up, the comparison of the two inferential models shows that the 2020 primary election model is better than the 2016 primary election model, with a higher R square value and F-statistic value, however, both two models are not significant enough.

**Predictive Models:**

First, we build the logistic regression model for both the 2016 and 2020 dataset and obtain the full model score of 0.6875 for both, while the 2020 model have a higher mean of y\_test (0.75) compared to 2016 (0.625).

Second, we conduct the feature importance with logistic regression for each feature, the result is as the following:

Model score with internet access only: 2016 (0.375). 2020 (0.437).

Model score with poverty only: 2016 (0.375). 2020 (0.375).

Model score with winning margin only: 2016 (0.375). 2020 (0.75).

Model score with battle ground state data only: 2016 (0.375). 2020 (0.75).

Finally, we build the model with feature importance using random forest, the result is as the following:

Full Model score: 2016 (0.5). 2020 (0.75)

Without poverty & MR & BG data: 2016 (0.625). 2020 (0.815)

Without internet data: 2016 (0.4375). 2020 (0.375)

I first built a logistic regression model for both the 2016 and 2020 datasets and obtained a full model score of 0.6875 for both. However, the mean of y\_test was higher for the 2020 model (0.75) compared to 2016 (0.625).

Next, conducted feature importance with logistic regression for each feature, and found that the model score was highest for the winning margin and battle ground state data in both 2016 and 2020.

Finally, built the model with feature importance using random forest, and found that the full model score was higher for the 2020 model (0.75) compared to 2016 (0.5). Removing poverty, MR, and BG data increased the model score for both years, while removing internet data decreased the model score.

**Summery and future work.**

Previous research has shown that among all the features, internet access has the highest coefficient with voter turnout, however, this result may not be significant enough. R square value only around 23% and 44%, does not have a high correlation.

The 2020 primary election model has performed better than the 2016 primary election overall, it has a higher F-statistic score, the main reason might be 2020 primary election has a higher internet access percentage which also has higher voter turnout, which might prove our hypothesis that higher level of political exposure in the 2020 election compared to the 2016 election, due to increased social media use.

In the feature importance with logistic regression for each feature of the predictive model, the model score was highest for the winning margin and battleground state data in both 2016 and 2020.

Finally, using the random forest to build the model with feature importance, it was discovered that the 2020 model's entire model score was greater than the 2016 models. While deleting internet data dropped the model score for both years, removing poverty, MR, and BG variables increased it. In particular, I gather the highest model score which is the 2020 primary election only with internet access data, the score is 81.25% which is the highest model score.

The research is still a lot of future work to make the model more comprehensive, including collecting more data which can increase the size of the dataset could improve the accuracy and reliability of the model. Consider collecting additional data points from more states or primary elections.

While internet access may have the highest coefficient with voter turnout, there may be other variables that also impact turnout. Consider exploring other variables such as age, education level, or political affiliation. And consider conducting further analysis to better understand the factors driving voter turnout.