STAT 6021 Final Report: Modeling Social Trust And Trump Approval

Group 3: Alex Bass(ujb3bu), Andy Ortiz(eao7r), Diana Morris(dmd8a), Grace Lyons(kat3ac)

Executive Summary

The Pew Research Center's American Trends Panel survey data of 2021 afforded us the opportunity to study attitudes and opinions associated with those Americans who are more or less trusting of others, and to understand what Trump supporters really look like with respect to its demographic information. We learned that the strongest predictors of social trust were the level of community attachment, feelings toward immigration (negatively associated), and age (older are more trusting). Trump supporters were well predicted by region, sex, community values, volunteer status, race, age, education, and political party. There were interactions between race and education, and between party and age.

Introduction

In 2016, Donald Trump was elected President of the United States, and - despite being banned from social media - continues to have a strong voice and influence in the current American narrative today. In recent years, many have said that politics has grown more divisive and polarized which could be a reason for Americans to be less trusting of their neighbors. In our project, we look at some of these ideas using survey data from the respected Pew Research Center. In our analysis, we explore two primary questions in America today:

1. What types of people are more and less trusting of others in American society?
2. What does a Trump supporter really look like according to reputable survey data?

In our report, we will first describe the data we used to answer these questions, share findings and insights from our exploratory data analysis, and then explain various models we used to answer these questions.

Data And Initial Cleaning

We used a data set provided by Pew Research Center's American Trends Panel. This is a probability based online panel of US adults. It was conducted between Feb 26, 2021 and Mar 11, 2021, and is the 32nd wave of this survey - though we only used this wave's data because it contained unique questions of our interest. Our original data set had 6,251 respondent observations and 145 fields which included all survey questions asked plus other metadata such as language, device type, etc.

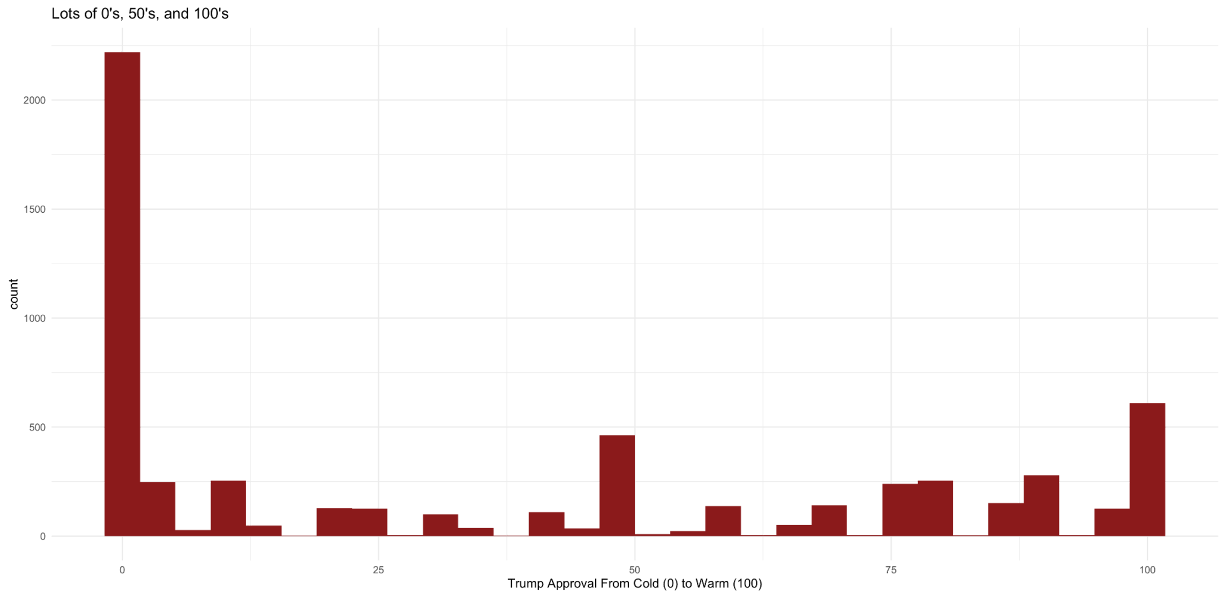
For our multiple linear regression modeling, we were only interested in the demographic variables as predictors and other questions as response variables in our model. We filtered the data to all of the demographic variables and the social trust and trump approval questions. From this point, we also realized that the survey questions were not required of respondents - meaning that many respondents skipped over questions they didn't want to answer leaving NA values in our dataframe. We also removed these respondents. After this initial filtering, we had 5,859 observations across 16 fields.

For our logistic regression modeling, we kept all of the variables except those for the presidential candidates. Since this larger set of variables contained missing data due to non-responses, filtering left 4,281 observations across 82 fields.

In both modeling cases we then converted the relevant variables to factors and collapsed several variables such as education, age, marital status, and others to have better categorical predictors in models. From this point, we were ready to dig into our exploratory data analysis.

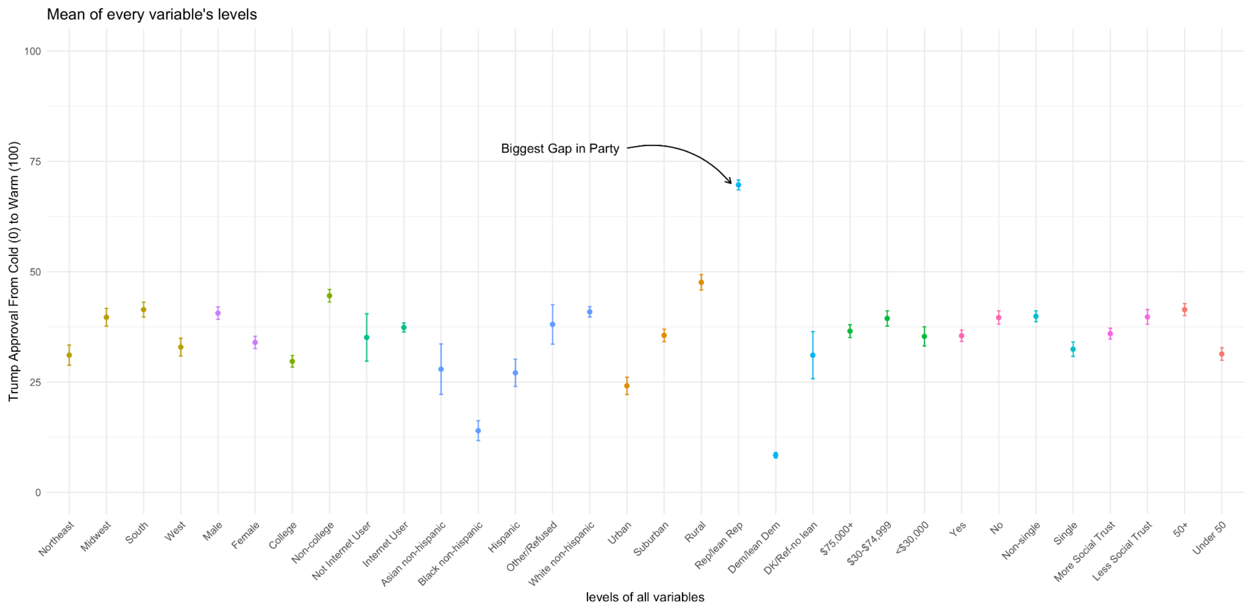
Exploratory Data Analysis

Looking first at the response variable (Trump Approval on a scale from 1 - 100), we notice a few things.

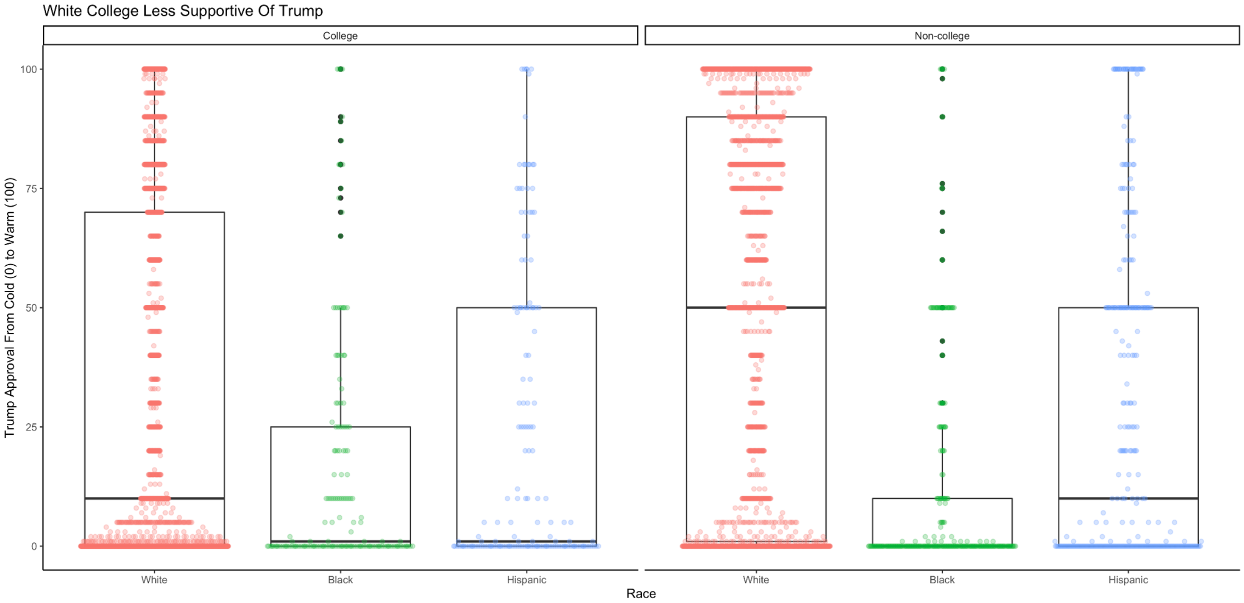


First of all, there are a lot more 0s (indicating coldness toward Trump) than anything else. Secondly, there are clearly some popular values such as 0, 50, 75, 100, and a few others. Even with these popular choices, the variable still has people filling in between making it a continuous variable. Because of this, we plan to apply multiple linear regression model to fit this data which we will elaborate on later.

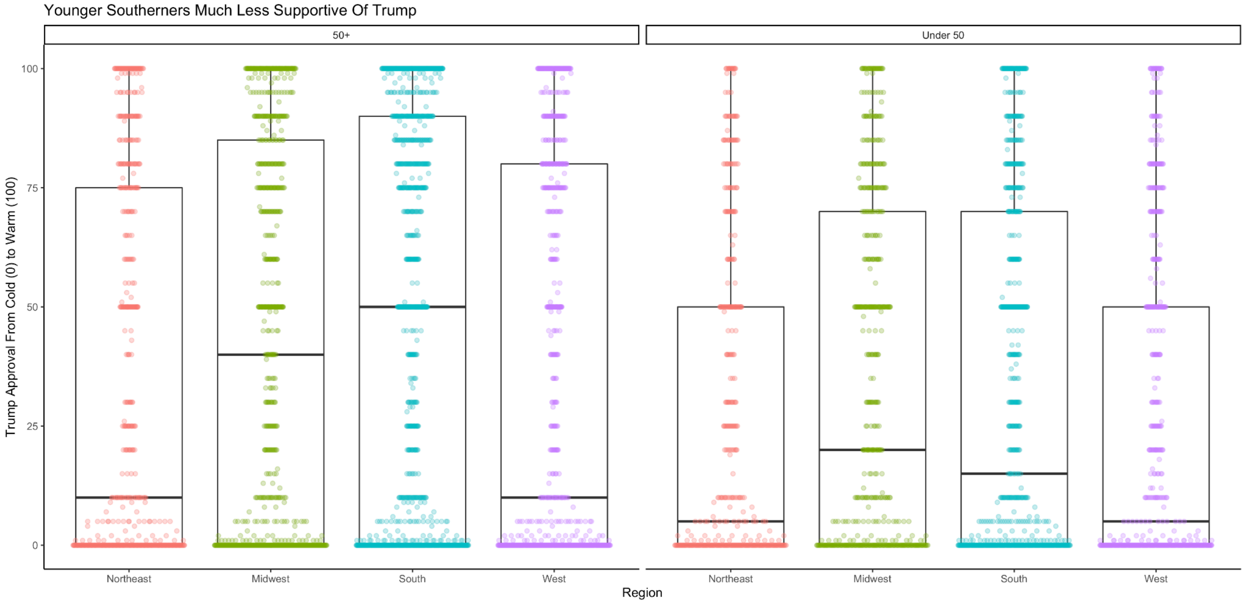
One challenge of our project was dealing with only categorical variables. Where normally, we would just look at a correllelogram or a correlation matrix, it didn't work as well with our data. Instead, we created the figure below which visualizes each of our demographic variables and the associated levels.



As is annotated, the largest difference we found was the respondent's political party affiliation. Other clear relationships existed between Trump approval and a respondent's age, education, race, and the type of community in which they resided (urban, suburban, or rural). From this point, we wanted to explore possible interactions we could use in our model between some of these variables.

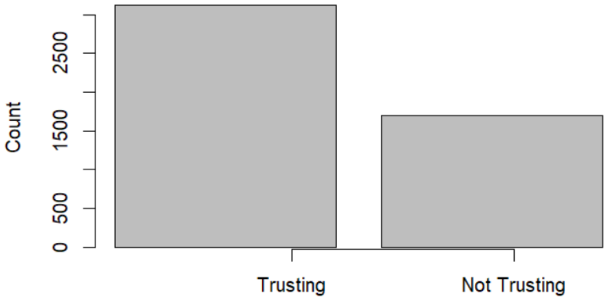


One strong possible interaction we found was between college attendance and race. The chart above overlays a boxplot with the beeswarm plot, so you can see both the density at a particular thermometer reading (0-100) as well as important benchmarks like the median and 25-75 percentiles. From this chart, white respondents who attended college are clearly much higher on trump thermometer ratings than white respondents who did not. From this chart, we decided to collapse our race variable from to white and non-white since that was the break we were most interested in.



Another possible interaction we found for our model is displayed in the chart above. We noticed that Trump Approval was higher among groups who were both age 50+ and lived in the midwest or the south. Younger people living in the south - while still more favorable to Trump than their younger counterparts in the northeast and west - were significantly less likely to approve of Trump than older southern/midwesterners. We explored and visualized other interactions as well, but did not find strong relationships (with the exception of Age and Party) and did not include them here for the sake of brevity. From this point, we felt confident to proceed with MLR model building.

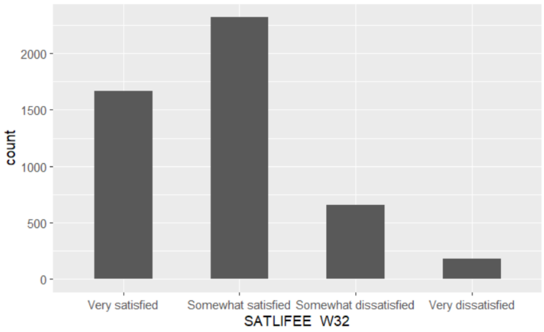
The logistic regression modeling used the binary response variable of "TRUST", which required the condensation of some of the social trust variable classes from 4 to 2, and 10 to 2. "Yes" for trusting was the reference class. The distribution of TRUST values was



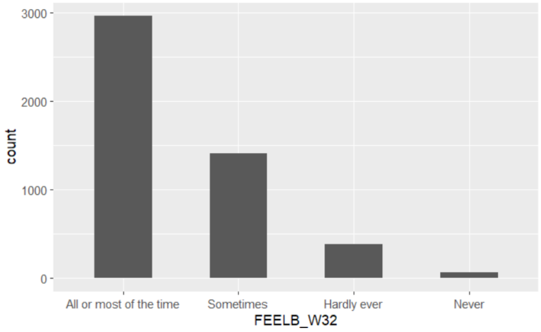
The trusting to not-trusting relationship was about 2 to 1. This imbalance may affect the results of our later predictions.

All of our predictor variables were categorical. Some of the plots of the distribution of classes within categorical predictor variables showed good variation among the classes:

“Are you satisfied with the quality of life in your community?”

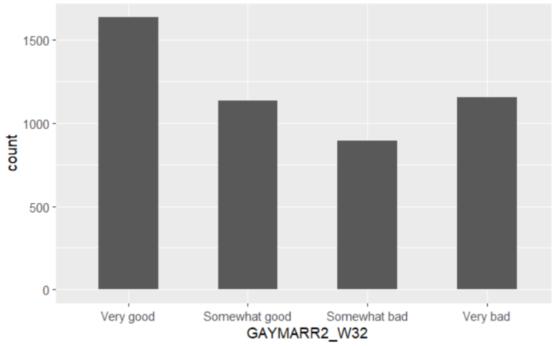


How often do you feel you have people you can turn to for support?

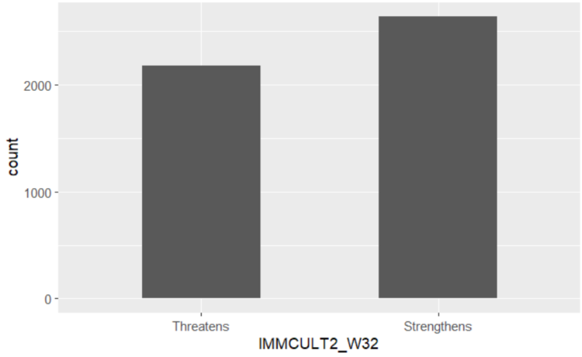


Other plots showed less variation among the classes:

“As you may know, same-sex marriage is now legal in the U.S. Do you think this is a good thing or a bad thing for our society?



“Which statement comes closer to your own views — even if neither is exactly right? The growing number of newcomers from other countries ____ American customs and values.”



Overall, we felt the distribution of the classes of the categorical variables were appropriate for our logistic regression.

Multiple Linear Regression

To understand the demographics of Trump supporters, we began our linear regression investigation by modeling Trump Approval against 12 categorical predictors as well as three interaction terms: region, sex, education, internet user, race, community, party, age, income, volunteer status, marital status, social trust, race and education interaction, party and age interaction, and community and age interaction. From the summary of the regression, we looked to drop some of the insignificant predictors to see if our model would still be significant. We decided to drop internet user, income, marital status, and the community and age interaction term. After doing a hypothesis test with the partial F test, we concluded that these predictors were insignificant and can be dropped from the model. We also ran a hypothesis test to drop social trust and region and found them both to be significant.

To test the validity of our model, we ran all possible regression tests as well as the automated search procedures. From all possible regression results we found the model with the best adjusted R squared, Mallow's CP, and BIC. The model for each criteria arrived at the same model, with race, community, party, age, volunteer status, race and education interaction, and party and age interaction being the best model. This model failed to find sex and social trust to be imortant predictors for the model. In our hypothesis investigation before we found both sex and social trust to be important predictors, thus we felt this model was not sufficient for predicting Trump approval.

After running all possible regressions, we ran forward selection, backward selection, and stepwise regression to see the models these procedures chose. The three procedures chose the same model, with region, sex, community, volunteer status, social trust, party and age interaction, and race and education interaction. This time the procedures picked the model that we had found during our own investigation. Further suggesting that the model we chose is the best model for predicting Trump approval.

With our model selected, we moved on to diagnostic testing of the model. We started by showing the VIFs for each coefficient. We found that region (south), race (white), education (non-college), age (under 50), race and education interaction (white non college), and party and age interaction (older republicans) all had relatively high VIFs with race and education interaction (white non college) being the highest. This makes sense with our model because we found that both the race and education interaction and the party and age interaction to be significant, thus we would expect to see these predictors with relatively high VIFs.

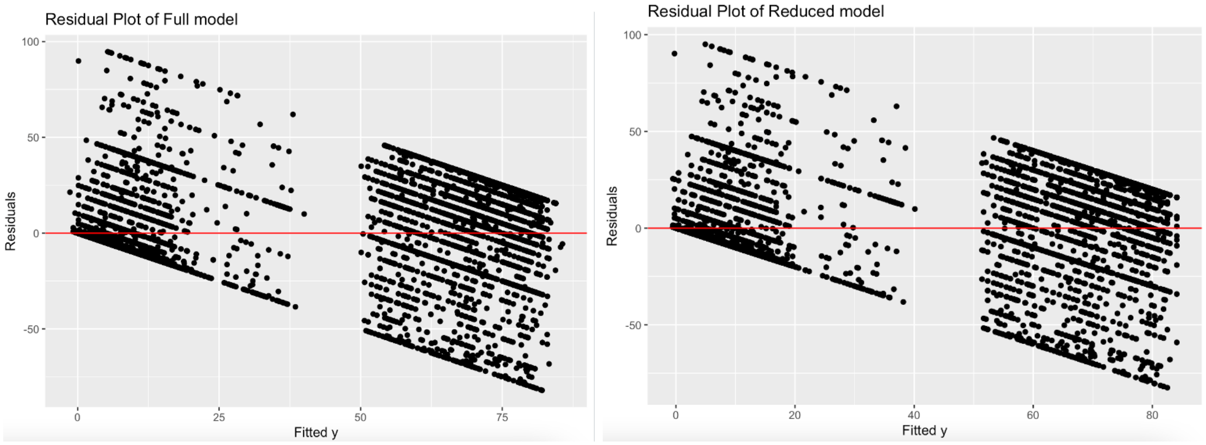
We also did an outlier analysis on our model. Cook's distance found no influential points while DFFITs and DFBETAs found a couple hundred influential points. We only ran the

DFBETAs for the intercept and the first two coefficients because we had 16 different coefficients. With the size of our data set, this is not surprising and though they are high numbers they represent only 5-7% of the overall population. We are comfortable keeping these terms in our model.

Cook's Distance	DDFITs	DFBETAs beta_0	DFBETAs beta_1	DFBETAs beta_2
None	321	321	444	416

Lastly, we plotted the residuals for our model on a residual plot and compared it to the residual plot of the full model. The two plots look very similar but the reduced residual plot has a slightly tighter fit to the x-axis. Though the residual plot is not perfect and does not tell us as much about our model as we had hoped, we are confident that our model can predict Trump approval fairly accurately after running the tests and the other diagnostics.

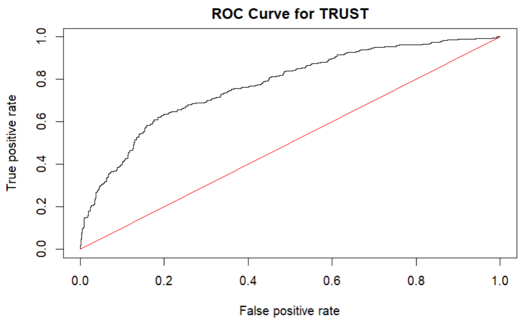
Full and Reduced Residual Plots



Logistic Regression of Social Trust

We wanted to understand factors associated with levels of social trust, as measured by the Pew Research Center's American Trends Panel. We split our 4,281 rows into 80% training and 20% testing dataframes. We began our search for a logistic regression model to predict the level of social trust by doing a logistic regression on the training set containing all of the 82 categorical variables, anticipating that many variables may not be significant predictors.

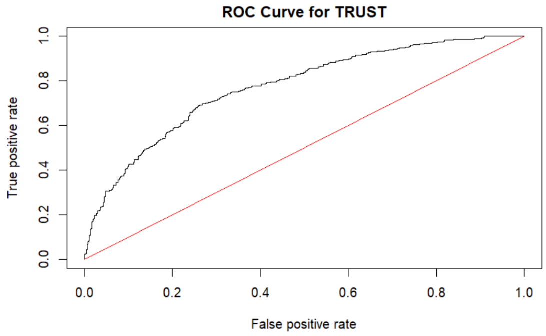
As expected, this full model had many non-significant predictors. Performing a prediction with this full set to establish the maximum possible predictive power of our data yielded the following:



AIC	**p-value**	**AUC**	**accuracy**	**FP**	**FN**	**cutoff**
4396	**0**	**0.77**	**0.75**	**0.13**	**0.47**	**0.5**

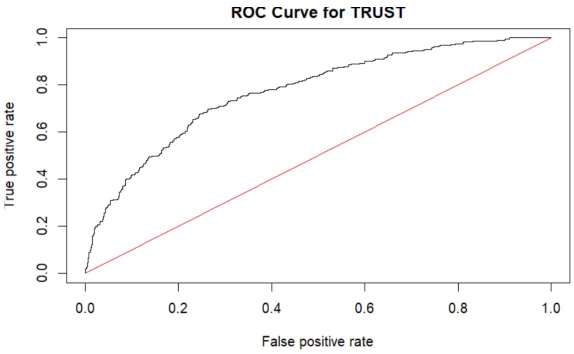
Not wanting to have such a large model containing unnecessary predictors, we removed the predictors with seemingly insignificant p-values, leaving a reduced model containing 20 predictors. After running a hypothesis test to determine if the reduced model was preferred to the full model, we found that the test was significant (p = 0.043), indicating that the full model does a better job of predicting social trust, but we judged that the vast simplification of the model justified this step.

The logistic regression output was still lengthy and is not shown here, but important information is as follows:



AIC	p-value	AUC	accuracy	FP	FN	cutoff
4284	0	0.77	0.74	0.14	0.5	0.5

A backwards stepwise procedure was then performed on this reduced model, with only one predictor being dropped, leaving 19:



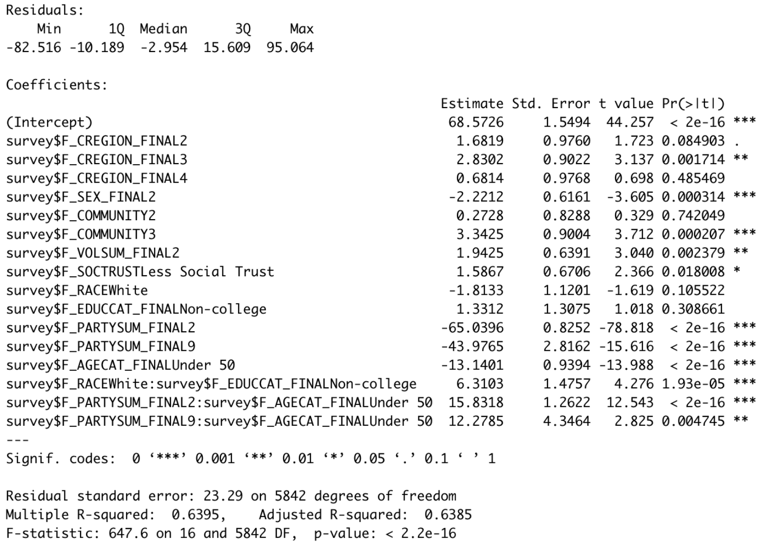
AIC	p-value	AUC	accuracy	FP	FN	cutoff
4280	0	0.77	0.74	0.14	0.5	0.5

At this point, all 19 predictors were significant and a hypothesis test revealed that the elimination of this last predictor did not affect the model. Considering the relatively poor predictive performance with a cutoff of 0.5, further reduction of predictors was likely to worsen the Accuracy, False Positive, and False Negative values.

Indeed, further attempts to simplify the model included eliminating demographic variables or alternatively eliminating attitudinal predictors. Hypothesis tests comparing each of these to the previous 19-predictor model yielded a p-value that rounded to zero, indicating that neither set could be dropped. It seems that attitudes cannot be extricated from demographics.

Limitations And Conclusion

Multiple linear regression



```
Residuals:
    Min      1Q  Median      3Q     Max 
-82.516 -10.189  -2.954  15.609  95.064 

Coefficients:
                                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                           68.5726     1.5494  44.257  < 2e-16 ***
survey$F_CREGION_FINAL2                                1.6819     0.9760   1.723 0.084903 .  
survey$F_CREGION_FINAL3                                2.8302     0.9022   3.137 0.001714 ** 
survey$F_CREGION_FINAL4                                0.6814     0.9768   0.698 0.485469    
survey$F_SEX_FINAL2                                   -2.2212     0.6161  -3.605 0.000314 ***
survey$F_COMMUNITY2                                    0.2728     0.8288   0.329 0.742049    
survey$F_COMMUNITY3                                    3.3425     0.9004   3.712 0.000207 ***
survey$F_VOLSUM_FINAL2                                 1.9425     0.6391   3.040 0.002379 ** 
survey$F_SOCTRUSTLess Social Trust                     1.5867     0.6706   2.366 0.018008 *  
survey$F_RACEWhite                                    -1.8133     1.1201  -1.619 0.105522    
survey$F_EDUCCAT_FINALNon-college                      1.3312     1.3075   1.018 0.308661    
survey$F_PARTYSUM_FINAL2                             -65.0396     0.8252 -78.818  < 2e-16 ***
survey$F_PARTYSUM_FINAL9                             -43.9765     2.8162 -15.616  < 2e-16 ***
survey$F_AGECAT_FINALUnder 50                        -13.1401     0.9394 -13.988  < 2e-16 ***
survey$F_RACEWhite:survey$F_EDUCCAT_FINALNon-college   6.3103     1.4757   4.276 1.93e-05 ***
survey$F_PARTYSUM_FINAL2:survey$F_AGECAT_FINALUnder 50  15.8318     1.2622  12.543  < 2e-16 ***
survey$F_PARTYSUM_FINAL9:survey$F_AGECAT_FINALUnder 50  12.2785     4.3464   2.825 0.004745 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.29 on 5842 degrees of freedom
Multiple R-squared:  0.6395,	Adjusted R-squared:  0.6385 
F-statistic: 647.6 on 16 and 5842 DF,  p-value: < 2.2e-16
```

There were several significant predictors we found in the model for our findings. Most substantially, we found the respondent’s political party affiliation to be the strongest predictor. Compared to Republican respondents, Democrats on average were 65 (out of 100) degrees lower on the feeling thermometer for Trump controlling for the other variables in the model. Also, age was a strong predictor with those under age 50 being an average of 13 degrees lower on the feeling thermometer towards Trump compared to those over 50. Additionally, we found two interactions to be significant that we noted in our EDA. Namely, those who are White and Older were more likely to rate Trump higher on the feeling thermometer, and those who are white AND didn't go to college rated Trump an average of 6 degrees higher. Other significant predictors include: Sex, Region, Community(urban, suburban, rural), and Level of social trust.

One limitation of this study is that this sample doesn't reflect the population of the United States because it is not truly a simple random sample. Respondents are drawn from a panel. A weighting variable was provided to approximate a truly random sample, but we did not use this variable in our analysis. In future study, one could test out a weighted least squared model to improve generalizability.

Logistic regression

Among the significant predictors in our final model, we noticed particularly strong associations between variables such as community attachment and level of social trust. The more attached a person is to their community, the stronger their feelings of trust. On the other hand, a negative view of immigrants was linked to less social trust.

More specifically, the difference in log odds of trust between those strongly attached to their communities (the reference class) and those totally unattached is 1.001, meaning that the odds of trust for those most invested in their communities is exp(1.001) = 2.72 times higher than for those with the lowest attachment ratings, holding other variables constant. Respondents with negative views of newcomers to the United States had odds of trust exp(-0.388) = 0.678 times lower than those who had positive views.

However, even the best model had mediocre predictive performance (accuracy 0.74) and a high false negative rate. Further work might include condensing the classes of the categorical predictors, where possible, to binary classes. Also, the only quantitative predictor variables contained in our data had been excluded in order to use them separately in our MLR model. We might want to reintroduce them for our LR model in the hope it would improve its predictive power.