SDS Final Project: Predict Counties' Political Leanings with Demographics Statistics

by Cheng Peng, Zhiyuan Wei, Erich Schwartz

Abstract:

Using the Data of 2016 & 2018 House elections and the social demographic statistics in 2761 counties across the United States, we construct predictive models of a county's political leaning based on the social, economic, racial compositions of the county's population. We discover that compared with model selection methodologies, tree models return better estimates in that the algorithm incorporates the interactions between different variables into the model. We also learn that the main predictors of political leanings vary significantly by region and is closely associated with the racial composition and the education level. Applying our random forest models to Texas, we also confirm the state's rapid shift to the left, and the trend is particularly prominent in north, east, and central Texas. Our estimates show that Texas would become a potential swing state in 2024.

Introduction:

American electoral politics has long intertwined with regional demographics. Groups of different income, race, and education level exhibit very different patterns in their electoral preferences. The Democratic electorate is known to feature minorities, women, and college-educated voters. The Republican electorate, in contrast, skews whiter and more rural.

The distinctions between the two parties grow more apparent in recent years as both parties swing to the more extreme sides of the political spectrum. Meanwhile, as a country of immigrants, America's demographics are always changing. States like Texas, Arizona, and Florida continue to trend Democratic, whereas Wisconsin, Michigan, and Pennsylvania gradually lean towards Republican, reshaping a presidential nominee's path to an electoral college victory.

While extensive research has been conducted to predict election outcomes, few have used social demographics to predict the political leaning of geographical units (like counties). Building a predictive model like this has many benefits. While election results are hard to predict (candidates' perceived likeability, for instance, is hard to measure quantitatively and polls are not always reliable in the age of the Internet), the demographic trends are relatively more apparent. Political parties and individuals can rely on models like these to best delegate resources and identify key constituencies a campaign should target.

This report uses data from MIT Election Lab and the Harvard Election Data Archive, which contains the demographical information of over 2700 U.S. counties as well as the results of four past elections (a.k.a. 2012, 2014, 2016, and 2018). We seek to build a predictive model of a county's political leaning based on its social demographics and analyze how each variable impacts a county's political stance. In addition, we use the population estimates from the Texas Demographic Center and apply our predictive model to determine the political leaning of 254 Texas counties in 2024.

Methods:

In this report, we use two datasets to conduct the analysis. The first one, election.csv,is derived from the MIT Election Lab and the Harvard Election Data Archive, which contains demographics and past election data at the county level. The demographics data features the total population, voting-age population, percentages of non-Hispanic whites, non-Hispanic blacks, and the Hispanic population, percentage of non-white population, percentage of foreign-born population, percentage of female population, percentage of population 29 years or under and percentage of 65 years and older, median household income, unemployment rates, percentage of the population without a high school diploma, percentage of the population without a college degree, percentage of the white population without a high school diploma, percentage of the population without a college degree, percentage of the rural population, and rural-urban continuum codes. The demographics data was collected between 2012 and 2016. The dataset also includes House, Senate, Gubernatorial and Presidential election outcomes from 2012 to 2018.

The second one, texas.csv, is obtained from the Texas Demographic Center, which contains estimates of the Texas population from 2020 to 2040 by age group, race/ethnic origin, and sex. It's important to note that different from the MIT statistics and the Texas Demographic Center classifies the population into five age groups: 0~18, 19~24, 25~44, 45~65, 65+. Therefore, we estimate the percentage of people 29 years or under by summing up the first two groups and 25~44 group divided by 4. We also use the data to evaluate the population proportion of different age groups, race, and sex. However, as social and economic data/estimates by county are unattainable, we assume the median income stays the same after adjusting for inflation. We also assume the education level and percentage of the foreign-born population remains the same in our estimates.

We apply two different mythologies and offer two prediction models. The first one is obtained by the model selection algorithm, which returns numerical estimates of each variable's slope coefficients. The dependent variable of interest is the Democratic vote share in the House election. In 2016, the Republicans were winning the generic ballad, whereas 2018 was a Democratic wave year. So, to offer more general estimates, we take the average of the two elections' outcome as our dependent variable. We choose the House election due to the large number of candidates running nationwide, so their personal attributes will have fewer impacts on the model. Senate, Gubernatorial and Presidential Elections may yield flawed estimates as few candidates were running and personal characteristics matter in U.S. elections. Our explanatory variables are all demographics variables from the MIT dataset.

We also use tree algorithms to build a second model, which is known to detect and incorporate interactions into the predictive model. We use the same dependent variables and explanatory variables from the first model and evaluate how the impacts of a variable like the white population percentage vary across different urban-rural settings and geographical regions.

Finally, we apply the two models to the state of Texas and estimate the political leanings of each county in 2024 using both models. The data is visualized with Google Maps and the information regarding the longitude and latitude of each Texas county. Low percentages of Democratic vote shares are denoted red while high percentages are denoted blue.

Results:

Model Selection Results (National)

Table 1: Regression Results

```
##
## Call:
## glm(formula = demvote ~ turnout18 + hispanic_pct + female_pct +
##     rural_pct + white_pct + lesshs_whites_pct + lesscollege_whites_pct +
```

```
##     median_hh_inc + age29andunder_pct + age65andolder_pct + clf_unemploy_pct +
##     lesshs_pct + lesscollege_pct + ruralurban + region, data = election)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.39713  -0.07860  -0.00933   0.06628   0.58219
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.916e+00  9.644e-02  19.863  < 2e-16 ***
## turnout18               -1.138e-01  2.900e-02  -3.925 8.88e-05 ***
## hispanic_pct             1.426e-03  3.608e-04   3.951 7.96e-05 ***
## female_pct               3.404e-03  1.156e-03   2.944 0.003264 **
## rural_pct                3.238e-04  1.391e-04   2.328 0.019971 *
## white_pct               -6.306e-03  2.722e-04 -23.164  < 2e-16 ***
## lesshs_whites_pct        7.771e-03  1.383e-03   5.621 2.09e-08 ***
## lesscollege_whites_pct -4.002e-03  1.307e-03  -3.063 0.002212 **
## median_hh_inc          -2.250e-06  3.609e-07  -6.236 5.18e-10 ***
## age29andunder_pct      -8.972e-03  9.384e-04  -9.561  < 2e-16 ***
## age65andolder_pct      -4.588e-03  1.204e-03  -3.810 0.000142 ***
## clf_unemploy_pct         5.932e-03  1.050e-03   5.648 1.79e-08 ***
## lesshs_pct              -9.517e-03  1.273e-03  -7.475 1.03e-13 ***
## lesscollege_pct         -4.615e-03  1.521e-03  -3.033 0.002441 **
## ruralurban2              3.582e-03  9.749e-03   0.367 0.713344
## ruralurban3             -1.266e-02  1.043e-02  -1.215 0.224581
## ruralurban4              4.420e-03  1.214e-02   0.364 0.715899
## ruralurban5              1.634e-03  1.654e-02   0.099 0.921285
## ruralurban6             -1.267e-02  1.031e-02  -1.229 0.219223
## ruralurban7             -3.974e-02  1.112e-02  -3.574 0.000357 ***
## ruralurban8             -1.503e-02  1.417e-02  -1.061 0.288925
## ruralurban9             -7.593e-02  1.330e-02  -5.710 1.25e-08 ***
## region2                 -1.768e-01  1.081e-02 -16.348  < 2e-16 ***
## region3                  6.573e-02  8.502e-03   7.731 1.48e-14 ***
## region4                 -6.022e-02  1.020e-02  -5.906 3.94e-09 ***
## region5                  6.573e-02  1.274e-02   5.159 2.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.01529867)
##
##     Null deviance: 90.281  on 2760  degrees of freedom
## Residual deviance: 41.842  on 2735  degrees of freedom
## AIC: -3677.7
##
## Number of Fisher Scoring iterations: 2
```

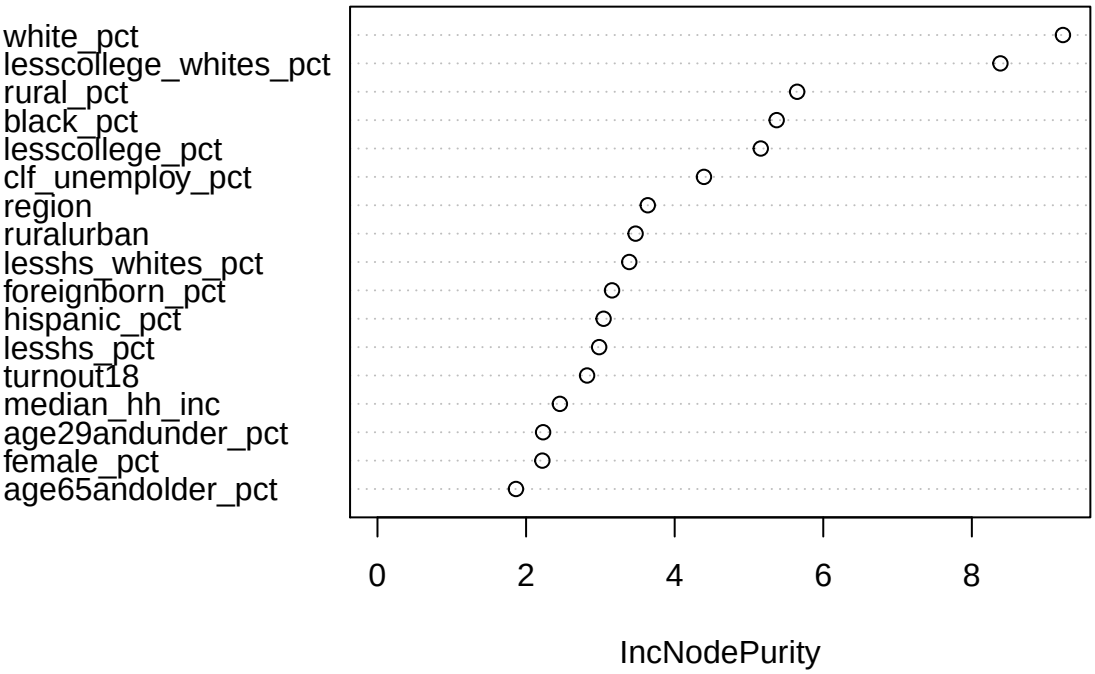
The Root Mean Square Error (in Sample) for the Model Selection is 0.123

Random Forest Results (National)

The Root Mean Square Error for the Tree Algorithm is 0.12

Clearly, the tree model yields more accurate predictions. Its out-of-sample performance is even better than the in-sample performance of the model returned by the model selection methodology.

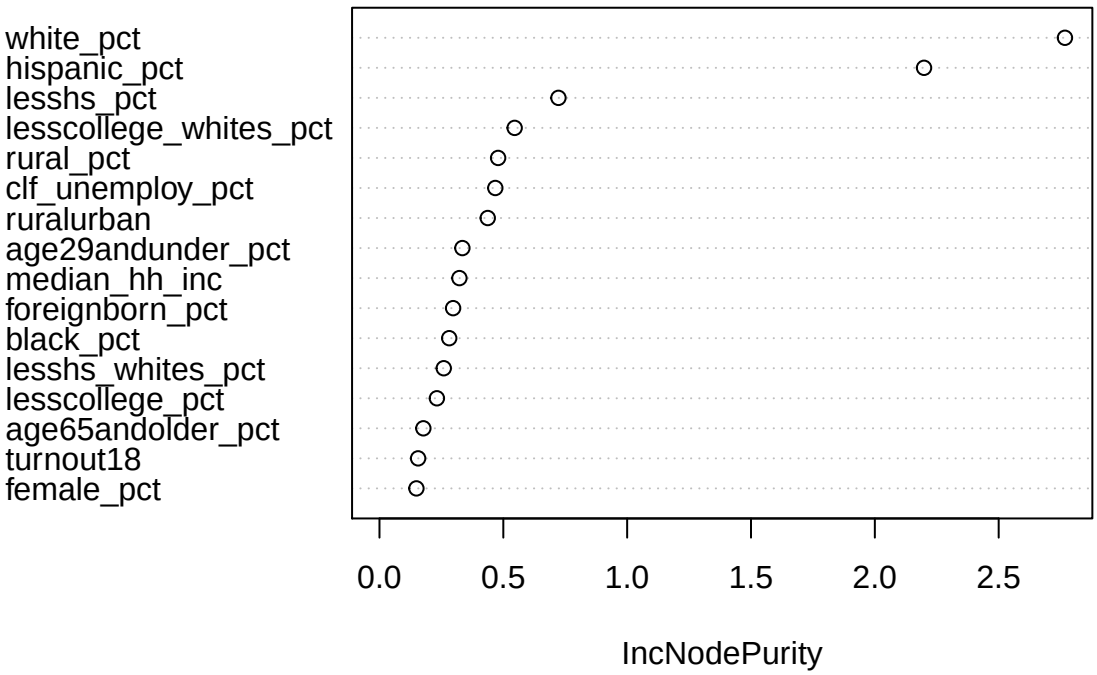
Figure 1: Variance Importance Plot (National)



The white population percentages, whites without a college degree population percentages, the percentages of people without a college degree, and the black population percentages are the strongest predictors of county-level political leanings nationwide.

Random Forest Results & Predictions (Texas)

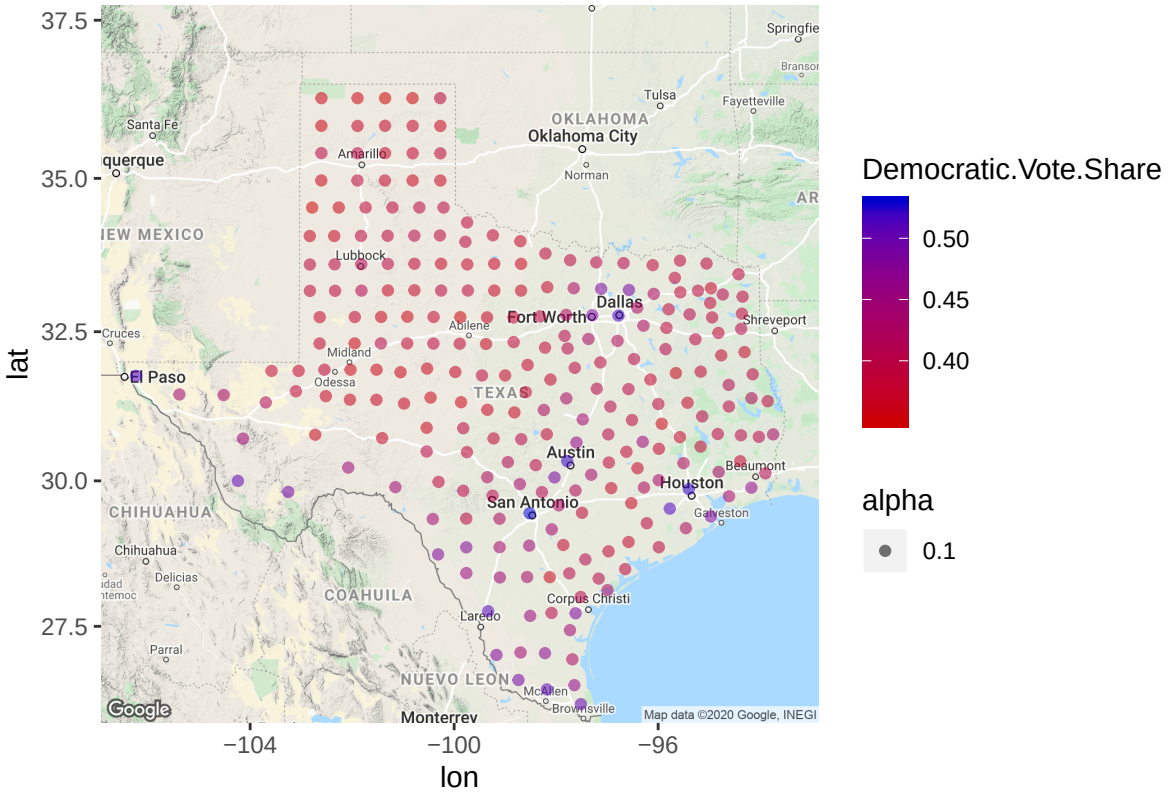
Figure 2: Variance Importance Plot (Texas)



The white population percentages and the hispanic population percentages are the two strongest predictors of county-level political leanings in Texas.

2024

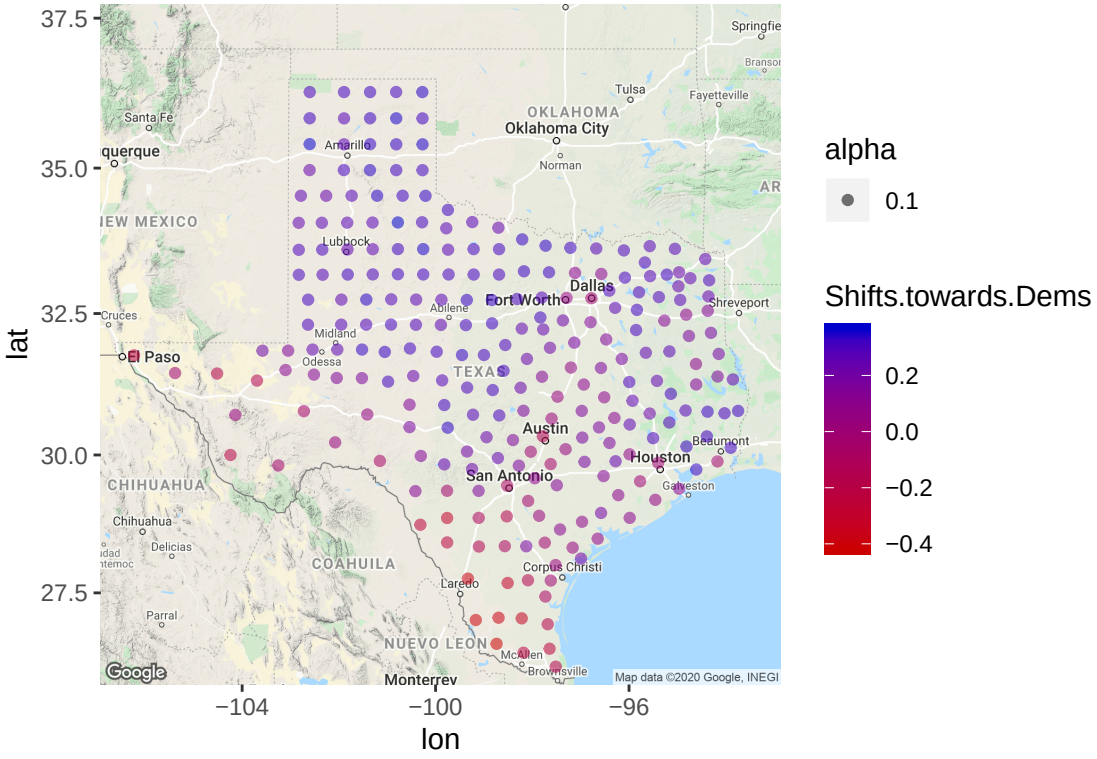
Figure 3: Democratic Vote Shares in 2024



Houston, Dallas, San Antonio, Austin, and counties near the Southern borders vote Democratic, while the rest of the state votes Republican in 2024.

The predicted statewide Democratic vote share in 2024 *is* 0.476.

Figure 4: Percentage shifts towards the Democratic Party in 2024



North, East, and Central Texas is trending Democratic.

Conclusion:

Building models with two different methodologies, we find it apparent that the random forest model yields better predictions (The out-of-sample RMSE for the random forest is 0.12, whereas the in-sample RMSE for the model returned from the model selection is 0.123). This demonstrates that variables of social demographics closely intertwine with one another. As evidenced by the Figure 7-9 in part B of the Appendix, in different regions, the effects of, for instance, the white population percentage is markedly different. The white population percentages have a more substantial impact on the Southwest and Southeast, with a larger white population than the rest of the country.

We also find from the two variance importance plots (Figure 1 & Figure 2) that the main predictors of the Democratic vote shares vary by region. On the national scale, the proportions of whites without a college degree within the entire population are the second most important predictor, followed by the percentage of the population without a college degree and the percentage of the black population. In contrast, the percentage of the Hispanic population is the second most critical predictor in Texas, followed by two variables associated with education levels. The differences make intuitive sense in that on the national scale, African Americans make up the second-largest electorate, whereas, in Texas, Hispanics/Latinos represent the second largest electorate. Also, education has long been considered a critical difference between the Democratic and the Republican voters. College-educated voters predominately vote Democratic. In 2018, 53 percent of college-educated white voters voted Democratic compared with 37 percent for those without a college degree. Also, compared to the United States as a whole, Texas may have lower average education levels. In both cases, the non-Hispanics white population percentage is the most important. This also demonstrates how regional demographic composition can shape the electoral outcome. We suspect the importance of the non-Hispanic white population is precisely due to the fact that non-Hispanic whites remain the dominant racial group and the largest electorate in Texas and America.

Our predictive model also showcases how demographic trends would impact Texas' political leanings. Figure 3 shows the predicted Democratic vote share in the 2024 House of Representatives based purely on social demographics. Though this is likely not the election outcome, as factors like incumbency and campaign finance strongly influence election results, it provides valuable information as to which county political operatives from both parties should target. Austin, San Antonio, Texas, Dallas-Fort Worth, and the counties near the southern border would continue to land in the Democratic column, while the rest of Texas still favors Republican ideologies.

In addition, from Figure 4, we learn that Texas, like the rest of the Sun Belt states, is rapidly trending blue. In most counties of the North, East, and Central Texas, Democrats are expected to get 0.1-0.2 percent of votes in Election 2024. The Democratic vote shares in four major metropolitan areas also appear to remain steady. More importantly, taking the weighted average of all counties' predicted Democratic votes share, we learn that in 2024, an estimated 48.3% votes in Texas would go to the Democratic Party, effectively making the GOP stronghold a battleground state in the years to come.

Finally, in part A of the Appendix (Figure 5 & 6), we use the model returned from the model selection algorithm to estimate Texas' political leanings in 2024. Though consistent with the tree model in suggesting the state's rapid shift to the left, the results return unreasonably high estimates for several Texas counties in the north. According to the model, some of the counties, now solidly Republican, would garner more than 60% of the vote in 2024, showing that the result is potentially problematic. It also reinforces our previous findings that tree models better estimate political leanings based on social demographics.

Appendix:

Part A: Model Selection Results for Texas in 2024

Figure 5: Democratic Vote Shares in 2024

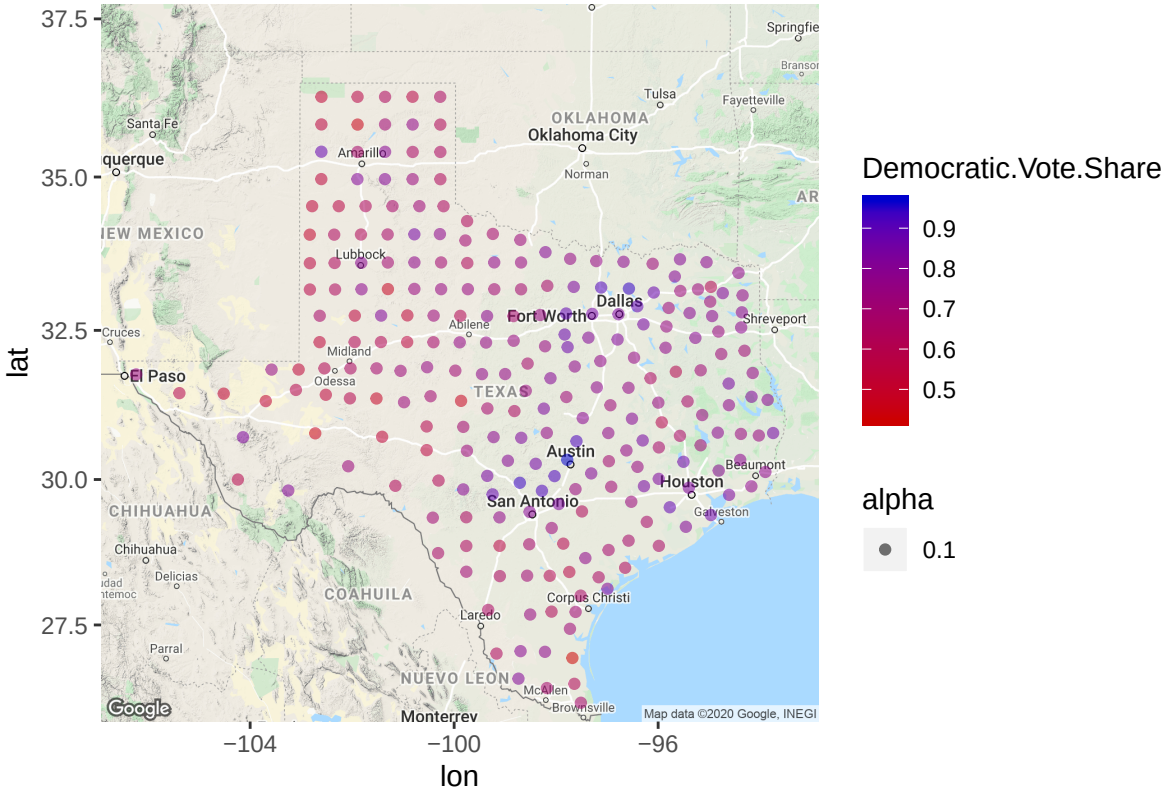
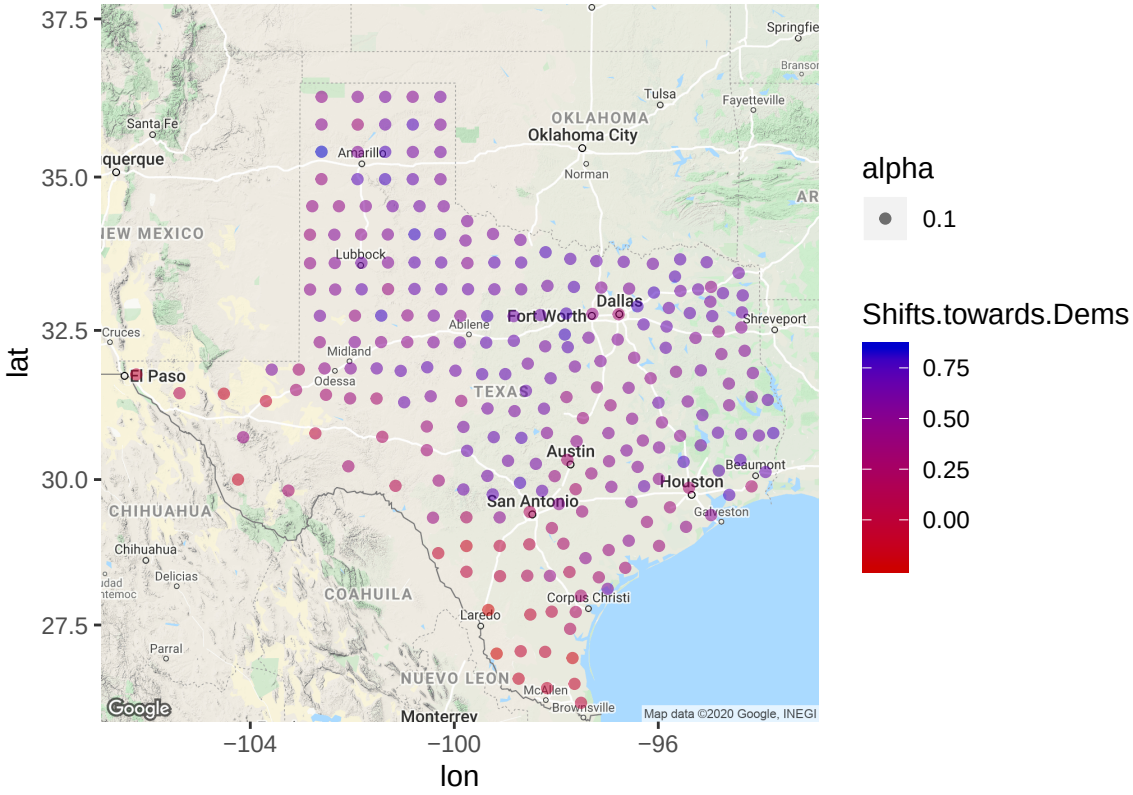
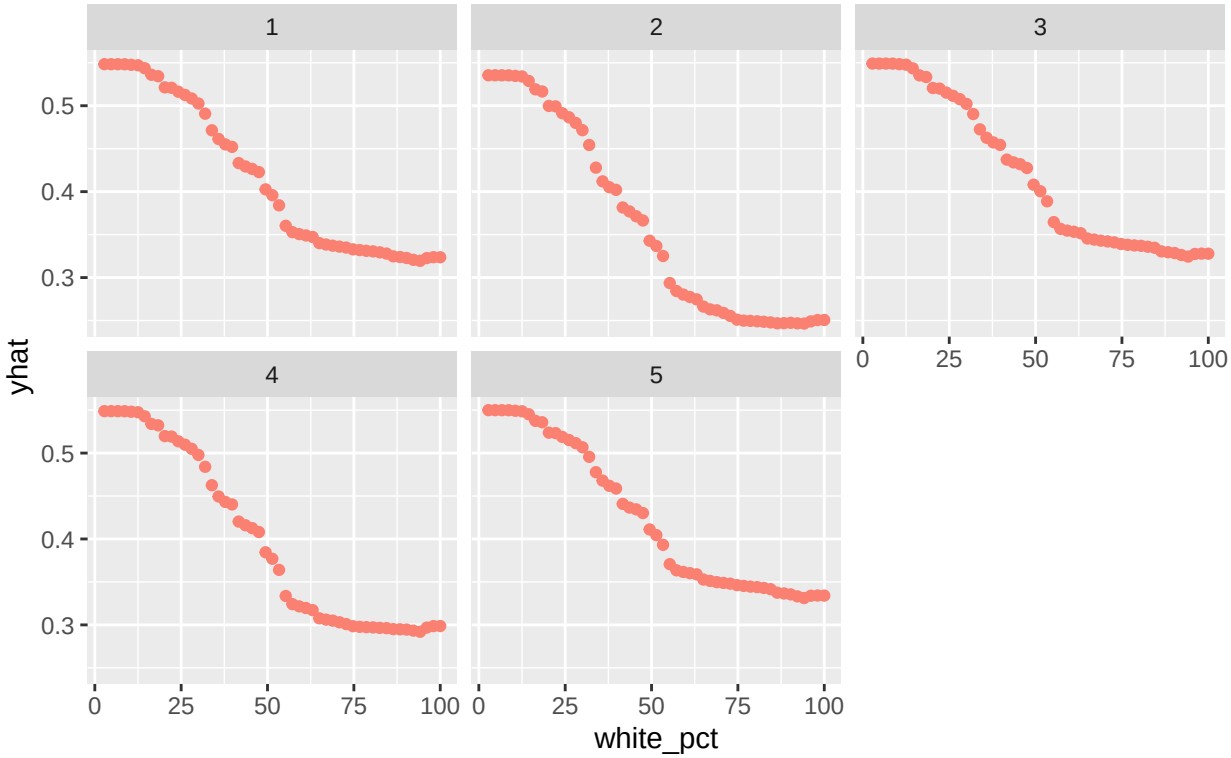


Figure 6: Percentage shifts towards the Democratic Party in 2024



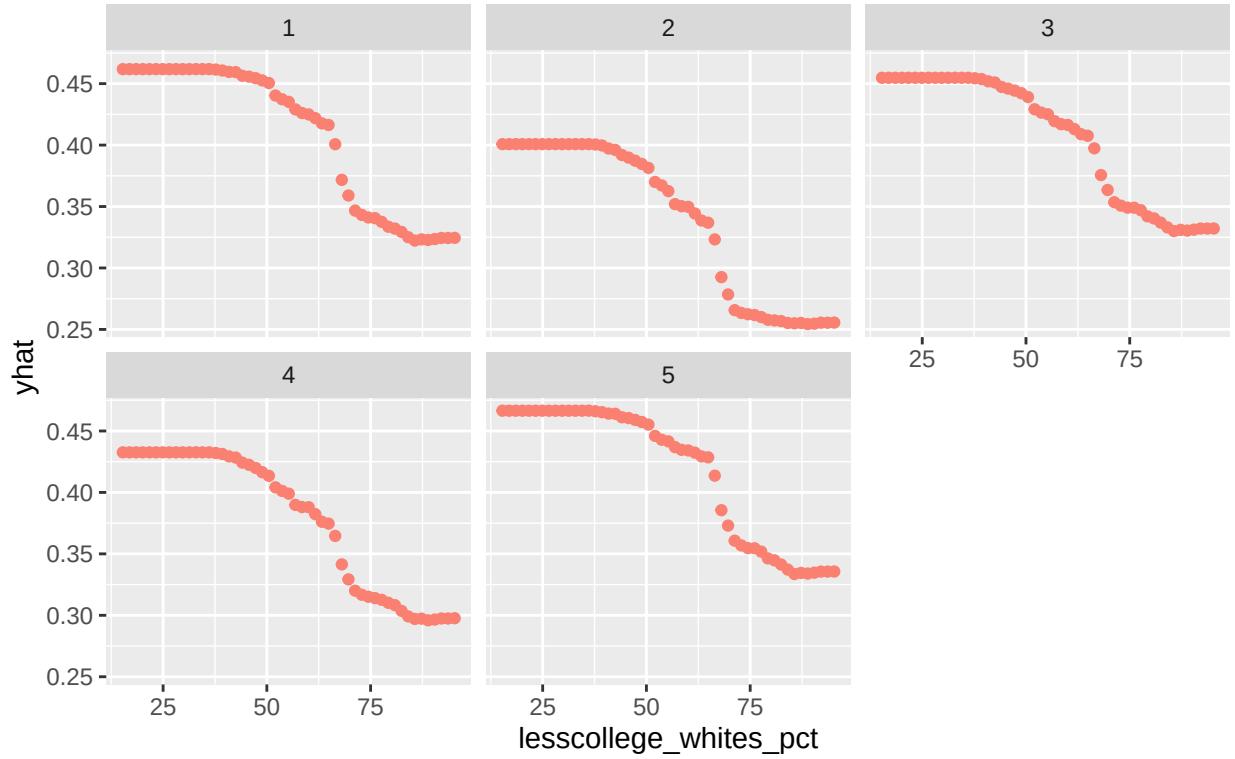
Part B: Examples of Interactions derived from the tree model

Figure 7: Effects of non−Hispanic whites population proportions across region



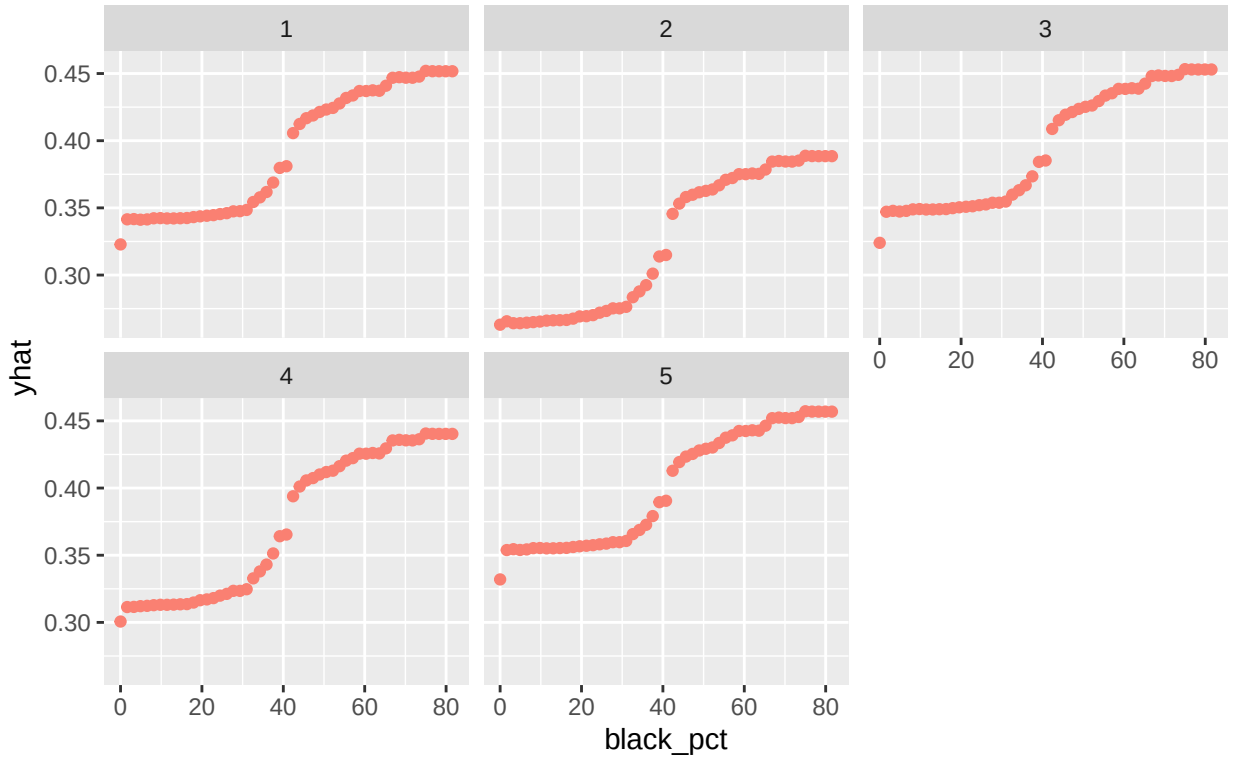
1−West, 2−Southwest, 3−Midwest, 4−Southeast, 5−Northeast

Figure 8: Effects of white population without a college degree proportions across r



1−West, 2−Southwest, 3−Midwest, 4−Southeast, 5−Northeast

Figure 9: Effects of non−Hispanic blacks populaton proportions across region



1−West, 2−Southwest, 3−Midwest, 4−Southeast, 5−Northeast