KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Influence of Demographics in Modern Politics

Joseph Stanton
P Johnson, PhD

Introduction

- Analyzing demographic components is crucial for understanding the dynamics of the American electorate and offers deeper insights into voting behaviors.
- In the post Civil Rights Era of the 1960s much attention was rightly focused on racial demographics, but now, education levels are becoming increasingly important with regards to choice of political party.
- This research analyzes demographic variables and their predictive ability for the 2020 presidential election at the county level.

Methods

- Demographic data were gathered through the United States Census American Community Survey (2015 - 2019). The survey is conducted yearly and aggregates estimates over a running 5-year period. Data are accessed using the R package tidycensus.
- The MIT Election and Data Science Lab curates data for election results (2020).
- The analysis focused on the county level because it offers the most consistent level of detail for election results. This contrasts with house districts and census tracts, which frequently have their boundaries redrawn.
- A broad range including 22 demographic variables considered to be potentially relevant resulted in complete case data for 3,113 (96%) of 3,243 county or county equivalents.
- No extreme correlation ($R^2 > .8$) or near 0 variance is present, thus no need for dimensionality reduction.
- An approximate 80/20 split was used for training and validation with all Georgia counties allocated to the testing data.
- A random forest model was tuned using these hyperparameters:
  - Split candidates: $2\sqrt{22} = 9$
  - Gini impurity was used as the split rule due to imbalanced data where approximately 15% of the data are Democrat counties.
    - $Gini = 1 - \sum_{i=1}^{k} P_i^2$
  - Minimum node size was trained on a range of 1 to 15.
  - Model preference was a split candidate of 9 and a minimum node size of 3.
- Models are validated using 10-fold cross-validation.

LinkedIn | GitHub

Table 1. Confusion Matrix

Prediction	Reference: Republican	Reference: Democrat
Republican	480 (TP)	13 (FP)
Democrat	15 (FN)	93 (TN)

Table 2. Measures of Validity

Measure	Value	Formula
Sensitivty	0.97	(TP)/(TP+FN)
Specificity	0.88	(TN)/(TN+FP)
PPV	0.97	(TP)/(TP+FP)
NPV	0.86	(TN)/(TN+FN)
Kappa	0.84	(P(Agree)-P(Chance)/ (1-P(Chance Agree))

Figure 1. ROC Curve

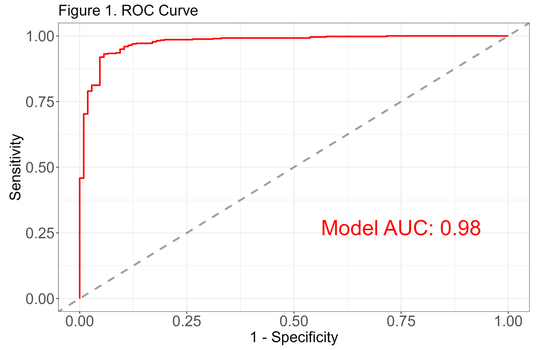


Figure 2. Scaled Variable Importance by Reduction in Gini Impurity

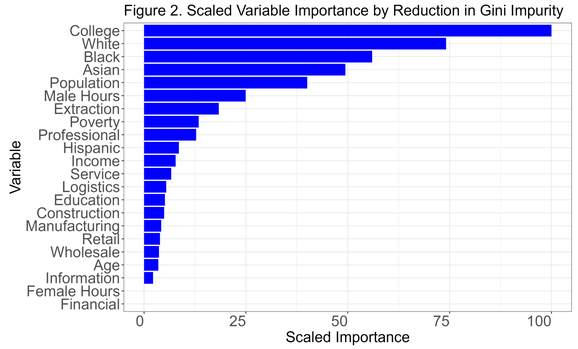


Figure 3. Probability of Voting Republican by Important Variables

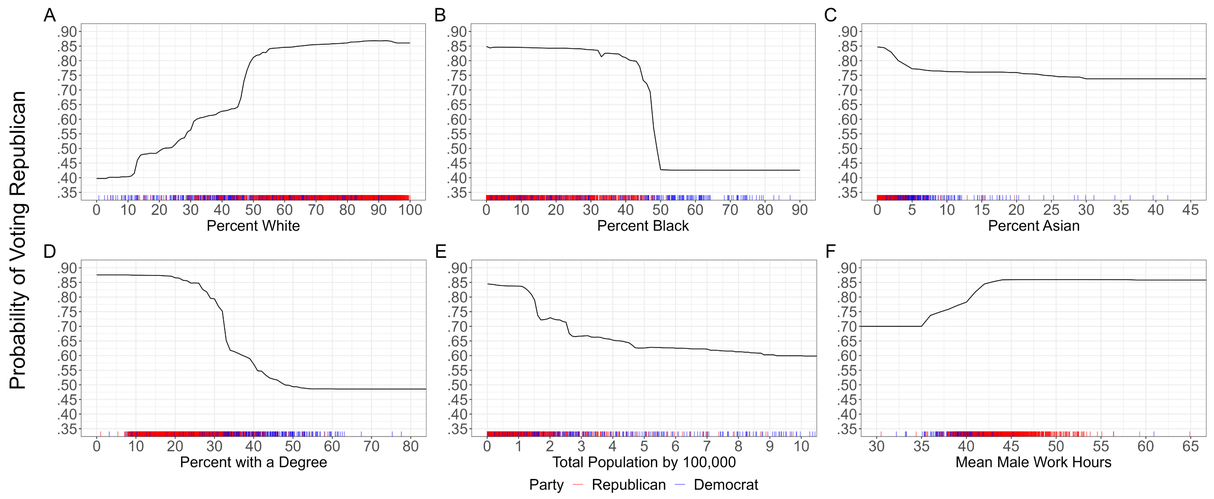
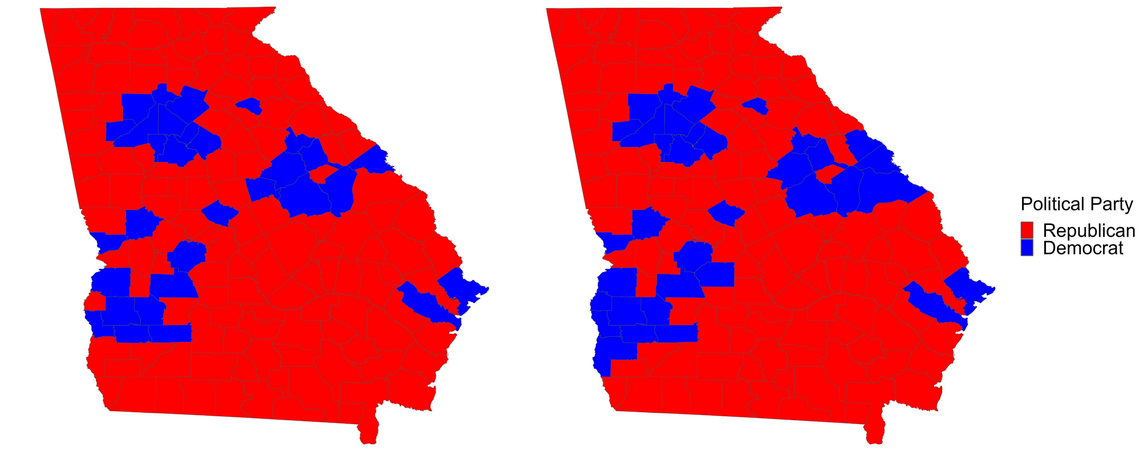


Figure 4. Map of the 2020 Presidential Election Showing Election Results (A) and Predicted Results (B)



Results

Confusion Matrix – 601 Counties (Table 1)

- The imbalance of counties by party can be seen with 495 Republican counties and 106 Democrat counties.
- 480 Republican and 93 Democrat counties are correctly identified.

Model Performance (Table 2, Figure 1):

- Of all Republican counties, the model correctly identifies 97% of them (Sensitivity).
- Of all Democrat counties, the model correctly identifies 88% of them (Specificity).
- A Kappa of .84 and AUC of .98 indicates the model has strong predictive ability beyond random chance.

Variable Importance (Figure 2)

- The variable identified as having the highest importance is the percentage of the county's population with a college degree.
- Racial demographics (White, Black, Asian) have 50-75% of the college degree variable's importance.
- County population is 40% as important.
- Male work hours are 25% as important.

Probabilities (Figure 3)

- As the percentage of white residents in a county increases, the probability of that county voting for the Republican party also increases (A).
- As the percentage of Black residents in a county increases, the probability of the county voting Republican decreases, with a significant decrease in this probability once the Black population exceeds 40% (B).
- The higher the proportion of the population that has at least a 4-year degree, the lower the probability that the county will vote Republican with a drastic drop in probability starting at 25% (D).

Georgia (Figure 4)

- 153 out of 159 counties are accurately classified.
- 1 Democrat county and 5 Republican counties are incorrectly classified.

Discussion

- The model demonstrates that demographic characteristics, especially education levels and race, influence voting behavior at the county level. The model accurately predicts Georgia counties' voting behavior even though no Georgia county was used to train the model.
- During the 2020 elections, Democrats received 60% of the vote from the 41% of the electorate who had at least a four-year college education.
- This research confirms trends in education levels and demonstrates the strength of the variable in predicting voter behavior.