Predicting the U.S. House of Representatives Election Outcome Using a Random Forest Model

Marc J. Alessi

Homework 1, Due: Friday, September 29

Collaborators: Kimberley, Bryam, Nicole, Josh

1. Motivation and Problem Statement

The United States House of Representatives (the House) is the lower chamber of the United States Congress, whose members are elected directly by equal-population districts across the country. The number of representatives grew steadily since the Constitution was adopted in 1789 but was capped at 435 members in 1939. The party that controls the House, Democrat or Republican, has the power to initiate the passage of federal legislation, impeach federal officials, and pass revenue bills that fund the government. The policy direction of the U.S., which has significant implications for the U.S. and global economies, is greatly affected by which party is in control of the House. This is especially true in an era of hyper-partisan politics; if the House is controlled by the opposing party of the President, for example during 2019-2021 under President Trump and during 2023-present under President Biden, the government in general cannot adopt major legislation that could positively (or negatively) affect Americans.

Given the implications of which party controls the House, election forecasts have become popular in recent years. *FiveThirtyEight*, an opinion poll analysis website with ties to the *New York Times*, began issuing House election predictions in the 1998 U.S. Midterm Elections. Their model issues predictions using a complex weighting system that considers polling, pollster historical accuracy, incumbency of the candidate, fundraising, the generic ballot, and past voting history in the state or district (https://fivethirtyeight.com/methodology/how-fivethirtyeights-house-and-senate-models-work/). Their classic model has a 95% accuracy rate 100 days before the election and a nearly 97% accuracy rate on Election Day. Despite this high accuracy rate, House election outcomes can still come as a surprise, for example, in 2020 Democrats were

expected to win 240 House seats, but barely maintained control of the House with 222 seats (218 is majority).

As an ardent follower of The Nates (Nate Silver, founder of *FiveThirtyEight*, and Nate Cohn, lead political analyst of *The New York Times*), believer of election forecasts (despite the soul-crushing year of 2016), and an individual interested in predicting complex non-linear systems (is humanity more complex than the atmosphere?), I attempted to create my own House prediction model using a Random Forest model. What is unique about this model is its use of not just historical election results in each district and polling, but also the use of demographic data including race, ethnicity, education level, poverty rate, and more. We demonstrate that our model has a nearly 96% accuracy rate with these predictors, and while it is not a perfect predictor of House elections, it may help guide political analysts in their forecasts of future elections. This could prove helpful in preparing the U.S. Government for potential conflicts between the President and House in the current era of hyper-partisan politics.

2. Method

To train, validate, and test our Random Forest, we use election data and outcomes from the 2014 and 2018 midterm elections, and the 2012, 2016, and 2020 general elections. We do not use election data prior to 2012 due to a general shift in American politics during the late 20th and early 21st centuries. For example, districts that primarily vote Republican today (e.g., deep south, majority White districts) voted strongly Democrat through the late 20th century and even into the start of the 21st century. After the 2010 midterms, the political landscape shifted to what it is today. If we incorporated election data prior to the 2010 midterms, the Random Forest might incorrectly predict Democrat for districts that are staunchly Republican today.

2.1. Data description

The demographic predictors (features) used to train the Random Forest are obtained from U.S. Census Bureau estimates for each Congressional District for the years 2012, 2014, 2016, 2018, and 2020. Specifically, our predictors are percent female; percent male; median age; percent White; percent Black; percent Asian; percent Native Hawaiian and Other Pacific Islander; percent two or more races; percent Hispanic or Latino (of any race); percent White (non-Hispanic); percent less than high school graduate; percent high school graduate, GED, or

alternative; percent some college or associate's degree; percent Bachelor's degree or higher; total foreign born population; percent over 5 years old that speakers English only at home; mean earnings (USD); percent below poverty level; and total number of renter-occupied housing units. Our reasoning for using these as predictors is because districts often vote based on the makeup of their population. For example, a predominantly Black district with less education strictly voted Democrat over the last decade. Typically, a predominantly White district with high education voted Democrat, while a White district with less education voted Republican, though these generalizations do not hold for all districts.

We also incorporate three more predictors: the generic polling average, polling for each district, and the 2019 Cook Partisan Voting Index (PVI). The generic polling average is the average of all polls conducted for the whole country that asks Americans whether they want Republicans or Democrats in control of the House. The generic polling average is the same for each district for each election year. For example, in 2014, the generic polling average was -2.4 Republican (negative denotes Republican), so all districts receive this value for this feature for that year. The polling for each district feature is 0 for most districts, but some pollsters conducted polls in districts in 2014 and onwards. Any poll conducted by the *New York Times*, Siena College, or Emerson in the last 4 weeks prior to the election was included in this feature. For example, Emerson conducted a poll in 2018 in Illinois' 14th district, which found the Democrat had a 5.5-point lead over the Republican challenger. This instance was changed from 0 to +5.5 (positive being Democrat). The 2019 Cook Partisan Voting Index (https://www.cookpolitical.com/cook-pvi/2022-partisan-voting-index) is an aggregation of how the district voted in previous elections. While the index is numerical, we set it to Republican or Democratic. We expect this to have the highest importance, since it provides information for how the district voted, Republican or Democrat, previously.

Our predictand is election outcome, Democrat or Republican.

2.2. Pre-processing and data preparation

The Cook PVI feature and Results were first converted to values (-1 for Republican and +1 for Democrat). In total, there are 2,175 data (435 districts * 5 election years). 25% of the data are used for testing, and the validation data are 25% of the original 75% training data. We use the 2019 Cook PVI as our baseline.

2.3. Model setup

Our model is a Random Forest Classifier. Originally, we set up the RF as a regressor (with a positive value being Democrat and a negative value being Republican), but this provided poor results against the baseline. A series of tests were run to identify which hyperparameter choices were most ideal for our model. We trained the RF model 30 times for 4 hyperparameters (number of trees, tree depth, node split, and leaf samples), meaning we trained the RF with the training data and then predicted the election outcome with the validation data 120 times with one hyperparameter adjusted each time (Figure 1). We selected our hyperparameters based on the highest accuracy and f1 values from these tests. The final values for our hyperparameters were:

num_trees = 24

tree_depth = 10

node_split = 2

leaf_samples = 1

RAND_STATE = 42.

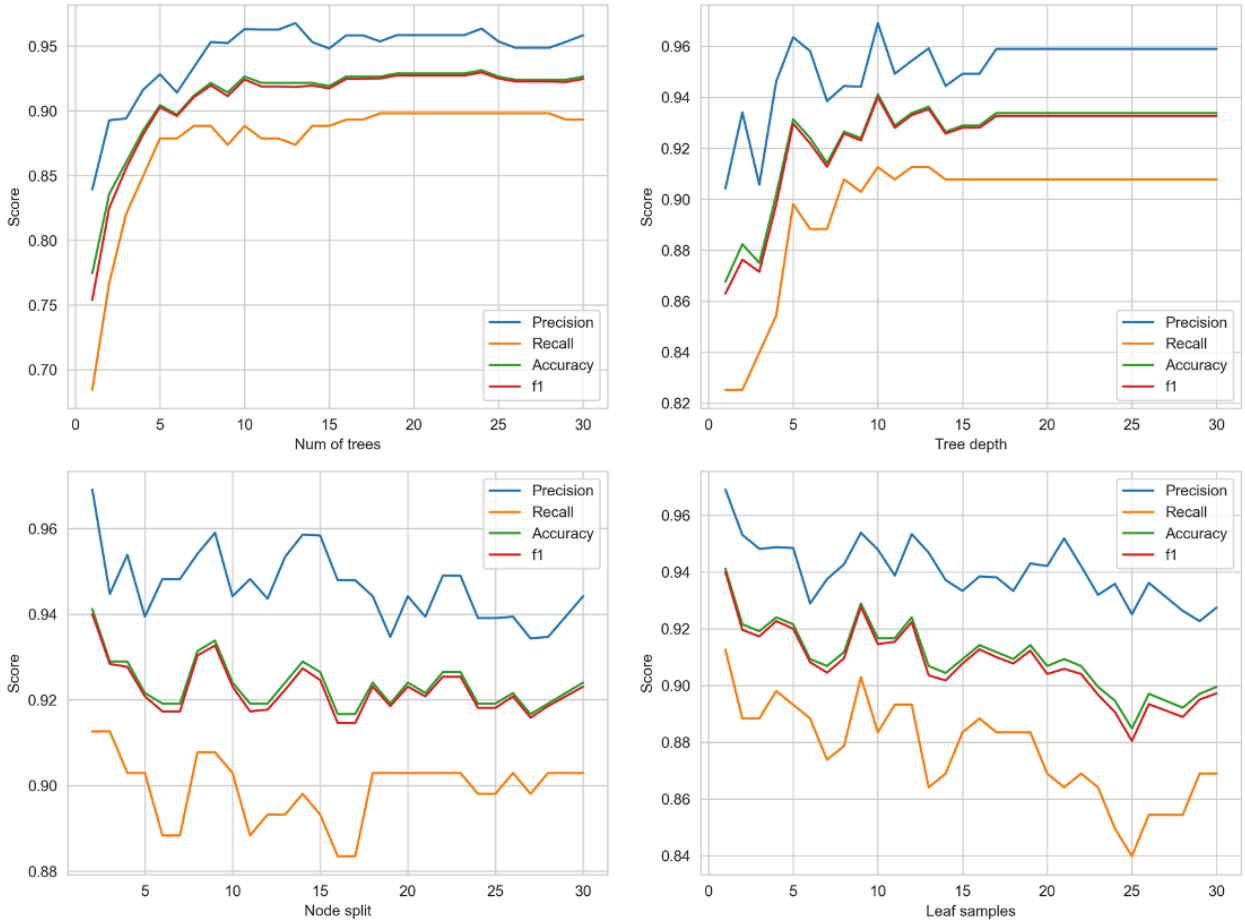

Figure 1. Precision, recall, accuracy, and f1 metrics for adjusting hyperparameters: number of trees (top left), tree depth (top right), node split (bottom left), and leaf samples (bottom right).

Though a tree depth of 10 is extensive, we found visualizing the trees to be quite interesting (not shown; a depth of 10 is too large to put into a figure in this document, though I have the visualizations available upon request). Tree #0 first split its data based on a threshold for total foreign born population, while Tree #1 split its data based on a threshold for percent Hispanic or Latino in a district. Following these pathways in a tree is also exciting, especially for districts that vote opposite of their Cook PVI value. In Tree #1, a district that has a low Hispanic population, a positive Cook PVI (Republican value), a large foreign born population, and is highly educated votes Democrat (gini value of 0.0!), despite the fact its Cook PVI is positive/Republican.

3. Results

We found our model to perform well during validation and testing (Figure 2). Out of 545 predictions for testing, only 48 were incorrect, which is an accuracy of almost 92%. The most important features in the predictions were the 2019 Cook PVI, total number of renter occupied units, percent that speaks only English, and percent of population in the district that was White (non-Hispanic) (Figure 3). This result is also reflected in the permutation importances, although percent with some college or Associate's degree was also found to be important in making a prediction for this metric (Figure 4).

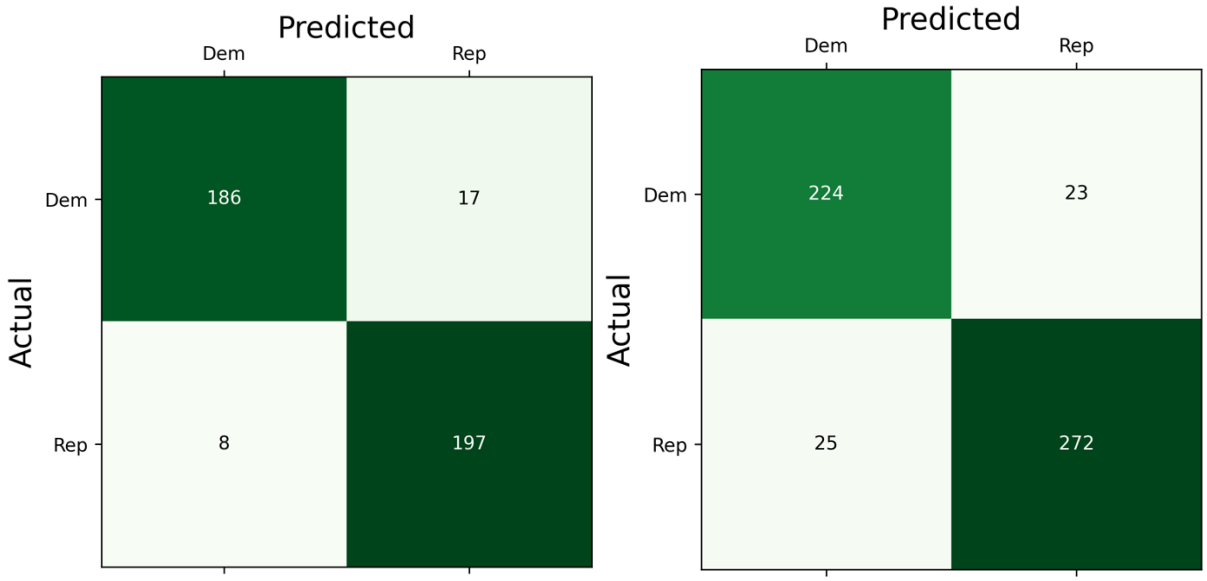

Figure 2. Confusion matrices for validation (left) and testing (right).

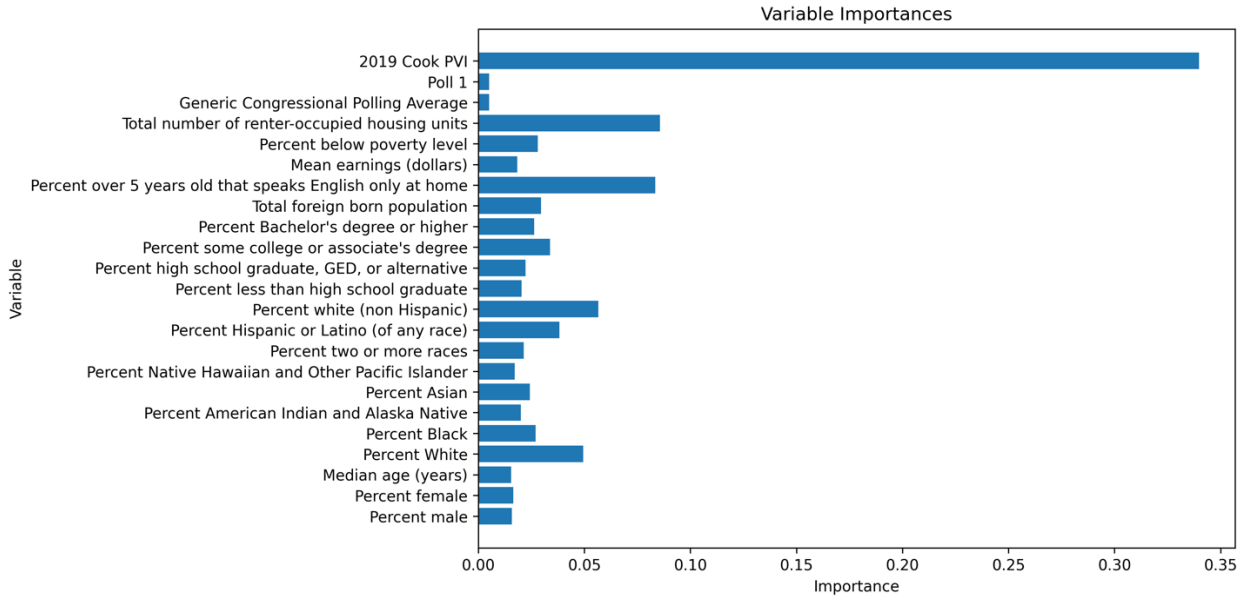

Figure 3. Feature importances.

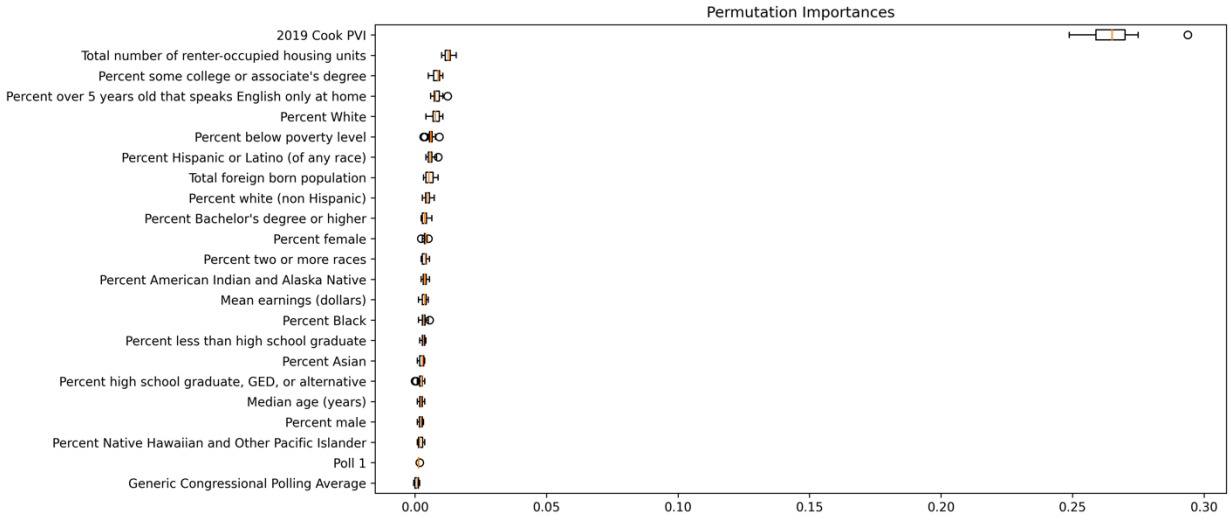

Figure 4. Permutation importances.

Given the steep drop in feature importance, we re-trained and ran the RF model with only the features that had a feature importance value greater than or equal to 0.03 in the original analysis. The confusion matrix and feature importance plots are similar to Figures 3 and 4 (Figure 5). The accuracy decreases only slightly, with 6 more Democratic seats incorrectly assigned to Republican. However, the model with less features correctly predicted 2 more Republican seats compared to the original RF model. Removing these less important features in future analyses makes the RF model less complex, which could prove helpful in interpretation of the results.

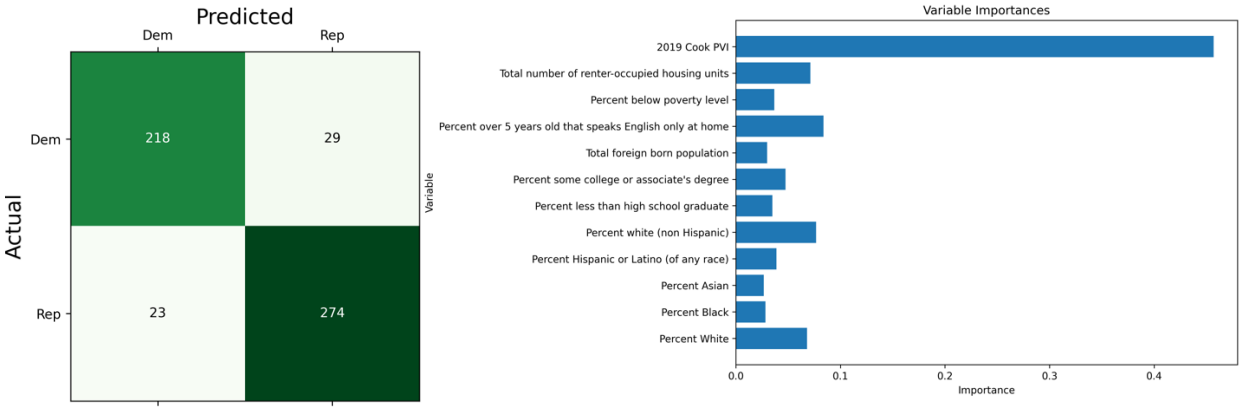

Figure 5. Confusion matrix and feature importance for the RF model trained with less features.

4. Discussion

An unanswered question is whether the model can correctly predict the outcome of a single U.S. House election. We ran the RF model again, but only using feature data for the 2020 General Election in order to predict the election outcome. The model in 2020 had an accuracy of nearly 96%, a score that is nearly identical to *FiveThirtyEight*'s model's aggregated accuracy for elections since 1998. We did not use *FiveThirtyEight*'s model as a baseline, since they primarily make probabilistic election forecasts rather than binary forecasts like our model. Instead, our baseline was the 2019 Cook PVI, which was also the feature with the highest importance in our model. Our RF model outperformed the Cook PVI for the 2020 General Election significantly (Figure 6). The Cook PVI accuracy was only 92% compared to our 96% accuracy. (Note that in order to use Cook PVI as the baseline, we had to remove districts with a PVI of "EVEN," or districts that are not typically Republican or Democrat in the historical record.) This is an exciting result! Initially, I expected the model to perform about the same as baseline, but the addition of demographic information seems to have allowed the model to predict the nuances of election-to-election changes in how districts vote.

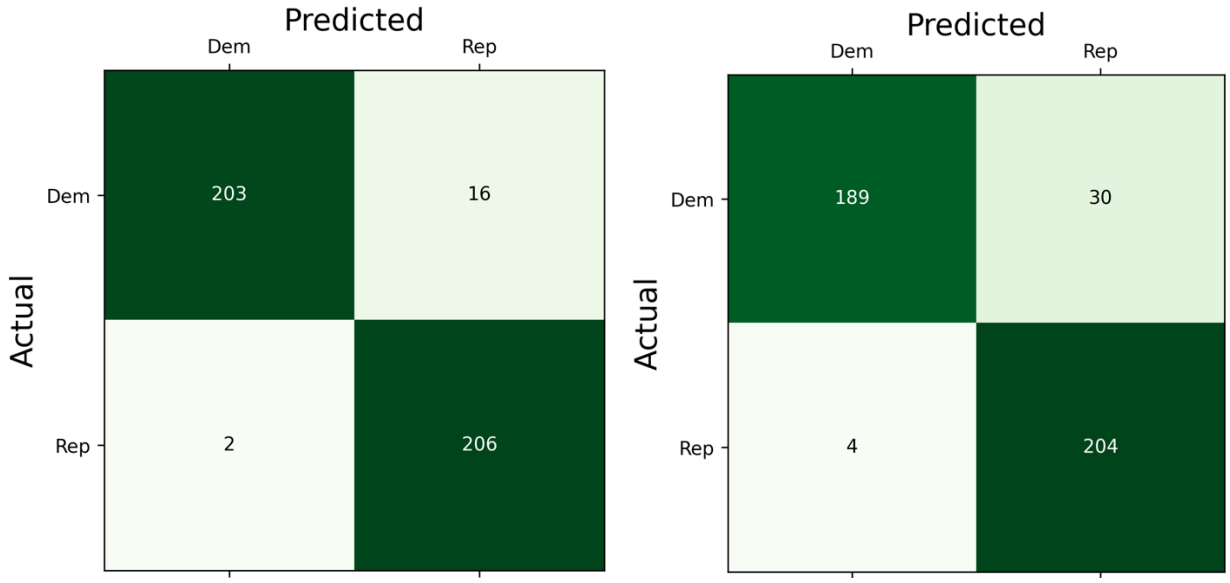

Figure 6. Confusion matrices for the 2020 General Election for our RF model (left) and the 2019 Cook PVI baseline (right).

An example of where demographics helped the model correctly predict the election outcome and where the Cook PVI (baseline) predicted the election outcome incorrectly is the example aforementioned in Tree #1: a district that has a low Hispanic population, a positive Cook PVI (Republican value), a large foreign-born population, and is highly educated votes Democrat. The relatively large foreign-born and highly educated population in the district helped the model choose Democrat as the outcome, since these two voting blocs increasingly vote more Democratic, rather than Republican. Another example for the 2020 General Election in which our model correctly predicted the outcome, Democrat, and the baseline did not is for Arizona's 1st Congressional district. This district covers Phoenix suburbs that vote Republican, and Native American Tribal land and the city of Flagstaff, AZ that vote Democrat. While the Cook PVI is Republican, the RF model selected Democrat for the outcome based on the heavily Native American population in the district and highly educated population in Flagstaff, which both vote Democrat.

The high accuracy, precision, and recall scores can be a bit misleading, but are still an impressive result. In the last 10 years, there have only been about 30-50 swing districts (a district that changes its vote between Democrat or Republican depending on the election year). While the RF model only incorrectly predicts the outcome in 18 districts for the 2018 election, this is still a large number that can decide which party controls the House. Interestingly, our RF model predicted that Republicans would win the majority in 2020 (only by 4 seats; see Figure 6), which perhaps could have balanced the significant error in *FiveThirtyEight*'s prediction of Democrats winning 240 seats (rather than the 222 they ended up winning). This election model is not perfect, but in combination with other election forecasting models, it could prove helpful.

We are under the impression that not much overfitting occurred in our model. The accuracies between the validation and testing data are both quite high and similar. If overfitting were the case, we would expect much lower accuracies for both validation and testing predictions, or just a lower accuracy for the testing predictions, depending on if the validation dataset was too small and did not sample enough House elections. On the other hand, one could argue that overfitting played a role in the 18 House seats that were incorrectly predicted by the model for the 2020 election. Why did the RF model predict so many Republican election wins than what occurred? Election forecasting is tricky because the political landscape is always changing. For example, during the 2018 midterm elections and as a result of Trump's win in

2016, Democrats gained significant ground among Whites in suburban areas (typically with higher education levels). Since the RF model trained on 2012, 2014, and 2016 elections before this shift in relatively educated, suburban, White districts occurred, it could have still predicted a suburban, White district would vote Republican instead of Democrat.

5. Conclusion

Overall, I found this project to be quite exciting. The skill in predicting a House election outcome by the RF is impressive. Incorporating this model into other well-established election forecasting models, such as *FiveThirtyEight*'s, could really help election forecasting efforts.

Something that was surprising to me while working with the RF model were the features that had low feature importance values. For example, a district with a large Black population (e.g., majority Black districts in cities and in the South) vote strictly Democrat, yet the feature importance for the percent Black category was lower than 0.03. Perhaps there are so few majority Black districts that the RF model found the feature to be unimportant in its predictions, and these districts all had a Democratic 2019 Cook PVI value, so the two features are not independent.

Github repository: https://github.com/mja244/ats780/tree/main