Our initial step was trimming the 13,353 rows in the dataset provided to remove noise from candidates who were not truly 'in the running', eg. lost in primaries or received less than 10% of the general election vote. This resulted in 5116 rows, and we further trimmed uncontested elections from the dataset as predicting the winners of these is trivial in 2018 and they would add noise to our model (eg. uncontested candidates may not try to raise money, lowering the usefulness of a 'Raised' or 'Spent' feature). We initially ran a random forest directly on these candidates, however found that it could classify multiple winners per district, so to resolve this we then separated them by 'district races', giving a dataframe of head-to-head races for each district.

Once we had our training dataset, we ranked the available features using domain knowledge, choosing PartyPreviousVoteShare, Incumbent, Raised and Spent. We selected PartyPreviousVoteShare because we posited most districts are 'sticky' and tend to vote for the same party year after year. The Incumbent feature was selected to consider the 'incumbency effect' - that incumbents generally have an advantage. Finally, "raised" and "spent" were metrics to help suggest how much resources candidates had available, and because more popular candidates are able to raise more finances. For our choice of machine learning method we used a Random Forest classifier with a Gini criterion, a max depth of 4 (gains after this were minimal) and 100 trees. We considered a few different classifiers such as multi-linear regression and decision tree but eliminated regression as we felt many features would have non-linear relationships to the winners (eg. Raised likely only increases likelihood of winning to an extent, then added utility drops off) and chose random forest over decision tree to reduce overfitting. Note, our forecast predicting a Republican majority is a result of our features being historical in nature (incumbency, PrevPartyVoteShare) and Republicans having House majority across our whole training set (2010-2016) - to solve this we would need a "polling" feature showing the recent blue wave in polls, which we will include for our final forecast.

The second part of the forecast was a prediction for the probability that each party would control the house after the election. Now, a naive approach would consider all the districts independent of each other, but Nate Silver has written extensively about how outcomes are tied. Thus, we chose to sample from a distribution of 435 Gaussians, with the means our probabilities that Dems won, and covariance .3 between districts, and .3 variance in the district. Ultimately, we'll use difference in location to pick covariances, but these were an estimate based on Nate's advice. For each sample, we gave the district to the Dem if they won with probability $>$ .5, and then they won that sample if they won more than 217 seats, then divided samples they won over total samples. Democrats have a 41.73% probability of winning a majority and Repubicans have a 58.27% probability.