Republicans or Democrats, who will win the 2018 and 2020 U.S house election?

The goal of this research is to conduct a classification model for my binary variable using some other variables such as years of experience and campaign finance of each party. Here we want our model to predict whether or not Democrat or Republican will win the 2018/2020 election. This research will focus on a dataset from the U.S house 2018 and 2020 for easy comparison between their wins and losses. This project is important because not many people get to see or understand why a party tends to win and why the other tends to lose and with this project being done, people can evaluate the candidates based on the amount of money they each spent on their campaign and how valuable it was to their country, society or district. Also, how much experience they had prior to this election.

Furthermore, it will be interesting for the voter to see the number of times a candidate has been elected which can help to predict what they will be able to offer for the next election they will be running for and determine their win based on who their opponent is. This is also important because it helps the voters to know which candidate or party to vote for in the next election.

Again, a candidate's years of experience as well as their campaign finance might matter for their next upcoming election they decide to run for which means they are likely to be chosen. However, in some cases candidates with more years of experience are likely to win the next election they run for compared to another candidate with less or no experience.

With that in mind, the focus of this research is to see how years of experience plays a role in a candidate chance of getting re-elected or elected compared to how much they spend for their campaign and I plan to use different models to predict whether or not years of experience plays a role in a candidate chance of getting re-elected compared to how much they spend on their

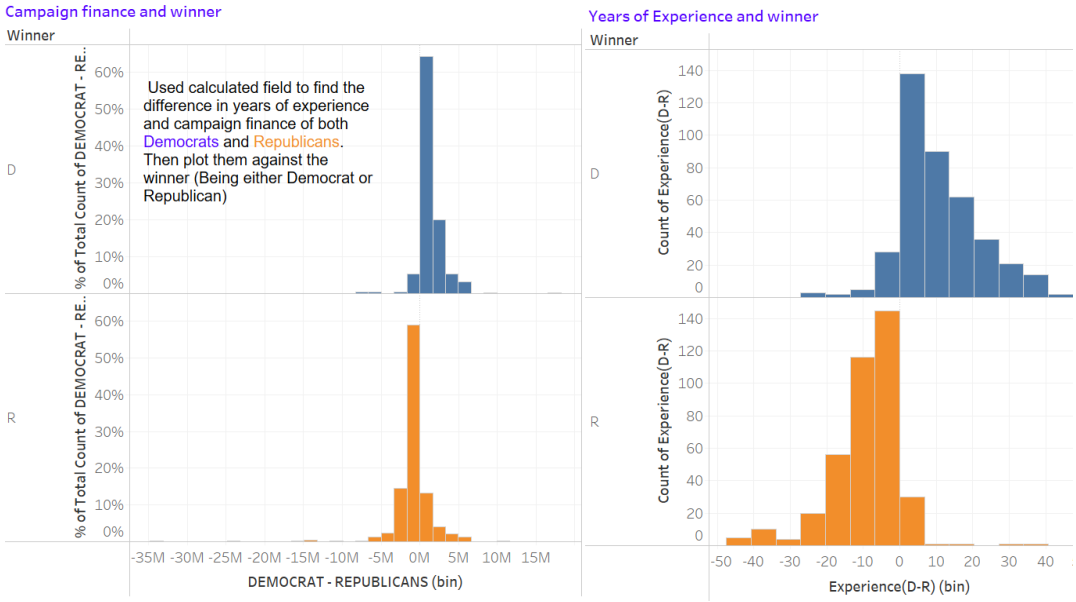
campaign. The two variables I will be using to predict my binary variables are years of experience and campaign finance of each candidate in the parties. The two variables were chosen and will be used because they are great at predicting whether or not a party will win or lose an election.

The dataset that I plan to use is from the U.S house elections for 1976-2020. It was collected using electoral votes from voters during that year but was stored in the MIT election data and science lab. The main focus of this project is on the 2018/2020 election and it is considered a reliable source because it provides enough data or information of each candidate using electoral votes. Hence, makes it easier to predict other variables that are not in the dataset with the available data.

Another reason why it is reliable is because the dataset was collected in an appropriate manner. In terms of collecting the number of votes, there could be potential errors in the number of votes for each candidate. By using these two predictor variables in my classification model, I hope to see how accurate these variables are at predicting who wins or loses the election.Moreso, to know how important it is for a candidate to have more years of experience prior to them running for election to determine if they will win or lose the election before having no prior experience.

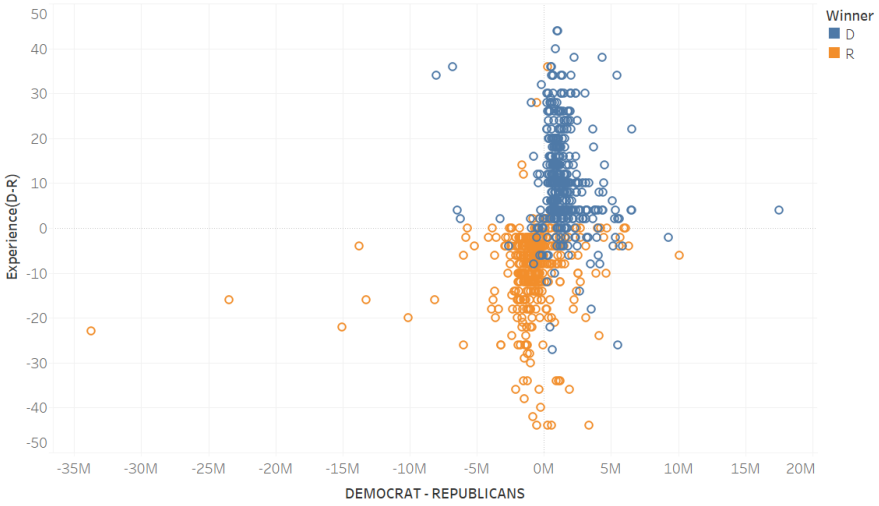
Before moving further with my classification model, I decided that it would be best to do some exploration with my data on tableau for better understanding of my research and to see if those two variables have good correlation. I decided to explore my data around the difference between years of experience (D-R) and campaign finance (D-R) of Democrats and Republicans. However, results show that as Republicans tend to have more years of experience they spend less money on campaigns and those that have less years of experience spends more money while

for Democrats, more experience means more money and less experience less money. The shape of the histogram below is unimodal and symmetric for both years of experience and campaign finance. Hence, this histogram below looks evenly distributed between republicans and democrats.



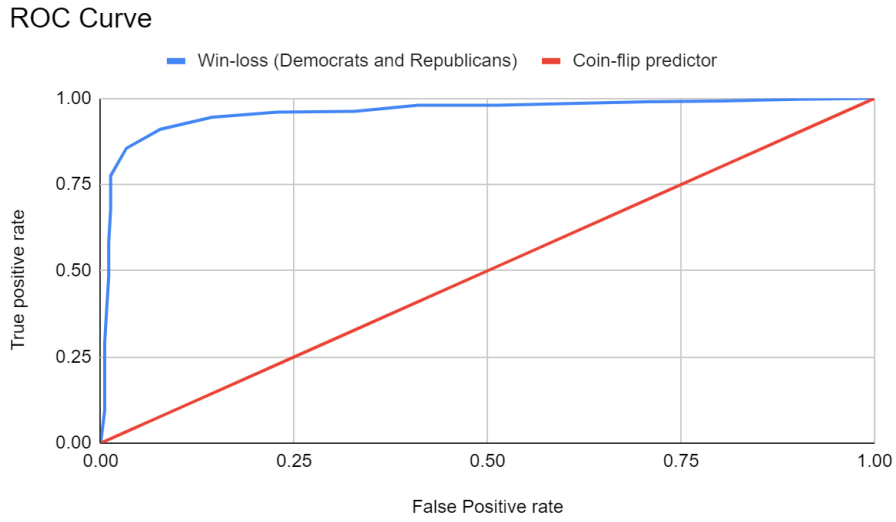
Furthermore, for better comparison and to see the relationships between my variables. I decided to create a scatter plot. I was able to plot a scatterplot of the difference in campaign finance and years of experience of both democrats and republicans to see the relationship between the two variables. The graph below allows us to see a clear relationship between my variables and the different parties. This shows a strong relationship and correlation between the two predictor variables which means that having spending more money does not guarantee a chance of getting elected but having more experience gives more chance of getting elected than the amount of money spent on a campaign. Most of the data are close to zero and in the negative too which explains that democrats have high chances of winning the election over republicans.

Scatterplot showing the relationship between Campaign finance and years of experience for each race/party.



DEMOCRAT - REPUBLICANS vs. Experience(D-R). Color shows details about Winner. The data is filtered on DEMOCRAT - REPUBLICANS, which keeps non-Null values only.

Having explored my data to see if there is any relationship between my binary variable and my two predictor variables. By using a classification model I can tell whether or not the two predictor variables are great predictors, if they are able to predict who will win the 2018/2020 election. Having inserted the data in google sheet, the results I got were more or less a surprising one because I did think democrats would win over republicans. Without changing the weight of the two variables, keeping them constant at one for the value of both weights, the variables were near perfect. That is, it did predict the outcome well because when I tried changing the weight of variable 2, the result thereof was changed to under the coin-flip predictor which is in the negative direction. However, changing the weight of the variable 1 which is campaign finance, makes the ROC curve a lot better. Therefore, the best weight (variable 2) for this model is 0.125, keeping the weight of the years of experience (variable 1) unchanged helps for a better combination of the result and producing a ROC curve like this below. Some else that could be done, could be changing or tweaking the weight of variable 1 being the years of experience, to a more better look for the ROC curve.



The best combination of weight created a better curve than the original weight of 1. The meaning of the ROC curve above illustrates that democrats have a high chance of winning than republicans because the model produces a much better outcome when democrats is predicted to win than when it is republicans, the curve goes underneath the coin-flip predictor, which in this case will not be a good classifier for our predicted outcome. However, according to the above model, this means that spending more money on a campaign is likely to help a candidate's chance of winning than years of experience. Although, this might not always be the case. Hence, there is a high likelihood of winning the election if the candidate spends more money on their campaign because by doing so, they are meeting the needs of their citizens or people, which then encourages them to vote for that candidate. Likewise, years of experience could also give a candidate a chance to win the election but like it was said, this might not always be the case.

Additionally, the ROC curve does a good job with illustrating the overall performance of the classification model alongside using different threshold values and the confusion matrix. Taking a close look at the threshold is a great way to see whether or not the model performed well, telling you that your prediction seems right. Having examined each point of the ROC

curve, the best threshold that predicted my binary variable well was between the 45 - 55 percentile but the best of all was the 45 percentile with a threshold value of -0.2327447331. Furthermore, I then constructed a confusion matrix with the threshold that had the best predictor, which can be seen below.

Confusion matrix	**Actual elected**	**Actual not elected**	Total
Predicted to be elected	379	22	401
Predicted to not be elected	56	334	390
Total	435	356	791

The reason why this threshold confusion matrix was chosen as one of the best is because the threshold values predicted the majority to be correct compared to the other thresholds. However, I believe that it is important to focus on democrats winning. This is because some republicans had few or no prior years of experience compared to that of democrats. That being said, I decided to choose that threshold and confusion matrix because it gives the best prediction.

To end this report, I will start by saying this model worked out really well when predicting which binary variable wins or loses. Moreso, there could be different ways or other improvements that can be done to make the model a better one to fit in with the result of the U.S house. The best way this model can perform well when predicting who wins or loses the election will be to suggest other variables that can help better predict the outcome variable, aside using the years of experience and campaign finance of each candidate or party. Such as candidate votes and which district the candidate represents. Furthermore, classification models can be very useful and can be a great predictor of our binary variable when it produces a good outcome but also

could have disadvantages due to the fact that an election vote could also go wrong during the election . Aside from that, it is done in a good manner.

Citation

Federal Election Commission *Browse candidates for House* (no date) *FEC.gov*. Available at: https://www.fec.gov/data/candidates/house/?election_year=2020&election_full=True&is_active_candidate=true

MIT Election Data and Science Lab (2022) *U.S. House 1976–2020*, *Harvard Dataverse*. Harvard Dataverse.https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FIG0UN2