Forecasting the 2024 US Presidential Election

Aleksander Bang-Larsen - 202107426

Introduction

In an effort to forecast who would win the 2024 US presidential election, we have built a bayesian forecasting model based on polls and sociodemographic factors. This kind of model is the de facto standard for data-driven predictions of the US presidential elections as it enables us to predict one-off events such as elections. A frequentist approach to this would be very limited as we have relatively few observations to include in this model. It is relevant for us to predict the election as a mean to understanding how the electorate responds to presidential campaigns and politics in general.

Model architecture

Our model uses a bayesian approach (Lock and Gelman 2010) to estimate the probability of winning for the Democratic candidate, which in this election was Vice President Harris. To do this, the model takes a prior belief of the candidate's probability of winning along with some evidence of the election to calculate the posterior which is from where we get our estimate. For our model, the prior given is the value calculated using a modified version of the time for change model presented by Abramowitz (2008). The original time for change model is theoretically defined and uses candidate incumbency, june approval ratings and second-quarter GDP growth of the election year to calculate a preliminary $p(Harris)$. This preliminary $p(Harris)$ is also the prior for our model. We gave the model a prior probability of Harris winning of 48.91. In our modified version we have exchanged the candidate incumbency in favor of consumer sentiment. We get the consumer sentiment data from University of Michigan (2024). We also added polling data on who the public deems to be the most capable at handling the issue that is most important to the respondent (Gallup 2024). Furthermore we added the national rate of new construction as a measurement of how much economic activity is 1) present and 2) visible to people. This data is from the Federal Reserve Bank of St. Louis and is calculated as

the rate of change since last year as that is almost analogous to how the common man perceives new construction - as more or less than recent years.

This preliminary prior is then held against estimated election outcomes based on polls. The polls are weighted such that the most recent polls have more influence on the final prediction, although all polls available will be used. The polls are collected by FiveThirtyEight as a part of their own election forecasting efforts.

Utilizing state similarities to get the most out of polls

To ensure that we have appropriate polls for all states we utilize a witihin-state correlation matrix to share polls between the states. This has the benefit of not requiring quite as many polls to reliably predict the election. Polls are after all somewhat of a scarce resource. This state correlation matrix consists of quite a few datapoints in our rendition. The basic model uses the percentage of white evangelicals, white working class, college educated, white percentage and the median age for the state. We have then added and filtered the percentage of black people in the state to *not* include the young black men aged 18-29 as they were more likely to be republican than other blacks in the 2020 election (Suggs 2024). We also added the percentage of hispanics not including cubans as they have shown to not be politically aligned within sync with the rest of the hispanic group (Krogstad 2020).

We have then added the median household income (based on 2023 data) on a state-level (Statista 2024) as well as the amount of urbanization in a state (US census 2020) and the percentage of religiously unaffiliated in the state (Pew 2024). All of these combined have shown to be a interesting map of correlation between the states and it showed us connections between states that are alike on politically significant factors. The median household income tells us how the state is doing in financial terms on a layman's level. The amount of urbanization helps us understand wether or not the state has mostly rural voters which usually are more republican. The religiously unaffiliated tells us how much the state is influenced by religion. This *could have* had an impact in this election with abortion being a major issue for one candidate, which also is an issue in christian areas.

By using so many different variables in our covariance matrix we have more datapoints to decide where to share the polls. In addition to this we have weighted three different kinds of factors in the state covariance matrix which lets the states correlate on the three variable subgroups by themselves to then be collapsed by a determined weight. We weighted previous voting for the state with around 25 points, the region in which the state is located by 10 points and the sociodemographic factors

mentioned above by 60 points. The remaining circa 5 points of weight is given to a correlation matrix that specifies 100% correlation between all states. We do this to ensure that some polling is shared between all states to grasp nationwide moves in the polls.

Overall $p(Harris)$

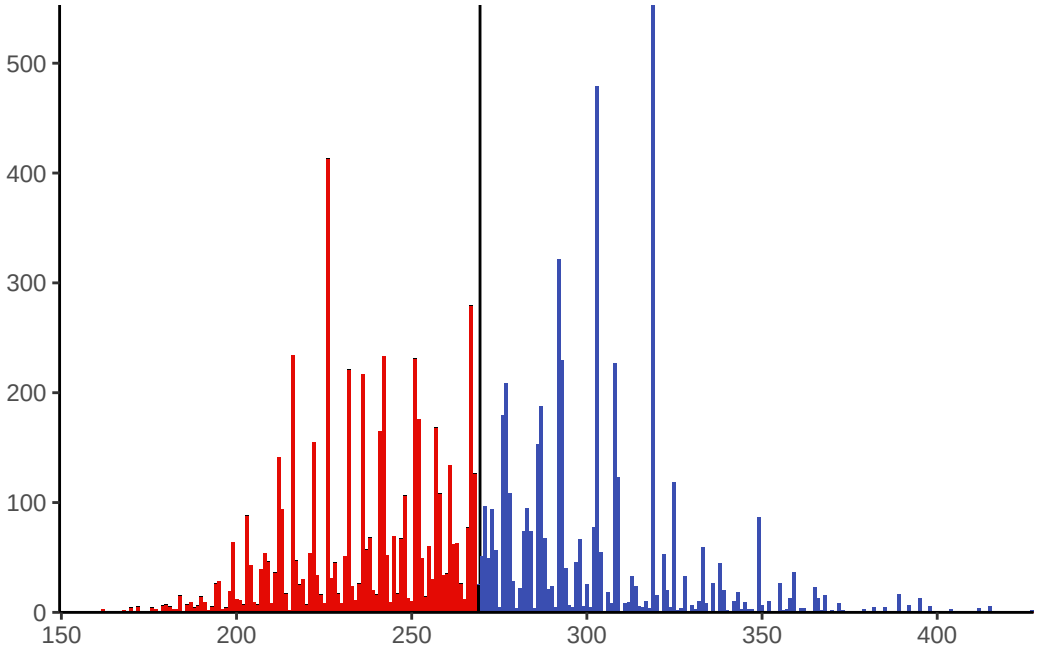
Our model performed rather well. In the sections below I will outline what went right and what went wrong.

Table 1: Probability of each election outcome

Democratic	No winner	Republican
0.4835	0.0025	0.514

The overall three most likely scenarios in our model was landslide victories for either Harris or Trump. In Figure 1 we see that the electoral college votes was pretty evenly distributed along the x-axis apart from the couple of high points. I deem it a success that our model was able to predict that either candidate would win in a landslide, exactly as it happened. The three most likely scenarios were either 319, 303 and 226 electoral college votes for Harris. In Table 1 we see that the overall $p(Harris)$ was a probability of 0.4835. This leaves a overall 0.514 probability for Trump winning.

Figure 1: Electoral college votes



Per state model performance

Figure 2: Predicted two-party democratic voteshare

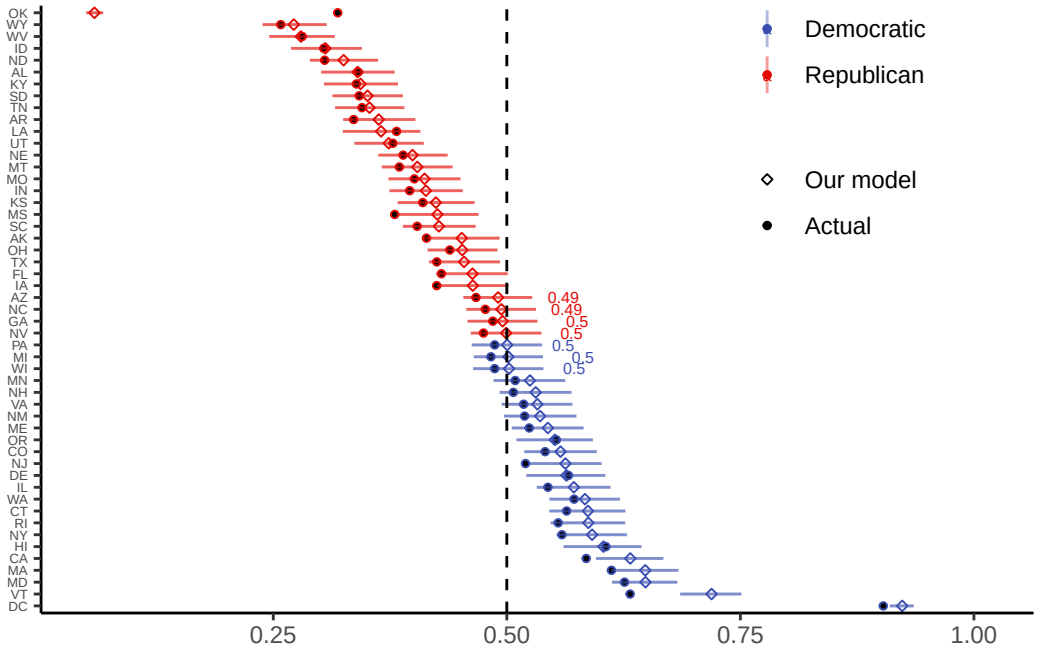
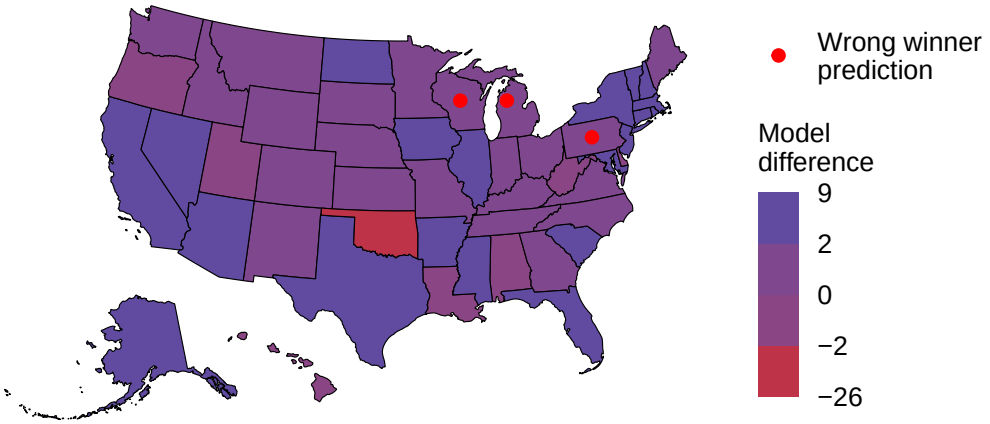


Figure 3: Difference from model to election outcome per state



In only three states, the wrong winner was predicted. These states are Michigan, Pennsylvania and Wisconsin.

In Figure 2 we see that all states except the three has the right prediction in terms of party while three of the swing states are actually wrong. On one side I find this somewhat disappointing as the three states that was predicted wrong was also three of the pivotal states. On the other side it is rather impressive given how close the polls were. Many of the polls in the close states were hovering around the 50 point mark. This exposes us to large changes in our model outcomes with even very small polling errors. I conclude that the errors of this model is due to either 1) polling error, e.g. nonignorable nonresponse from young men or 2) the model's sensitivity to the prior or 3) the state-correlation matrix.

In Figure 3 we see that most states are within a respectable +/- 3 points (purple colors) with some larger states being more than three points too democratic, such as Texas and California. Note that Oklahoma had a very low predicted democratic voteshare in our model (0.06), which is a right prediction in terms of electoral college votes, but was -0.26 points off in terms of getting the distribution right. This is the worst performance of the model, but it does not affect the end result as the difference between Oklahoma voting 30% democratic and 45% democratic is actually zero, in terms of electoral college votes at least.

Changes for next time

The things that went wrong in this model is also things that I would consider changing for the next presidential election. Specifically the sensitivity to the prior is something that would make the model better. For TFC-prediction i would exclude the construction measure. Also, I would reconsider the state correlation matrix to specify even more sociodemographic variables. Also an addition of media markets or social media usage could be interesting.