CS 760: Machine Learning Fall 2020

Climate Change and Partisan Politics

Authors: SOLLY PARENTI, YANDI WU, HYEJIN (JENNY) YEON

Abstract: Climate change has become an increasingly partisan political issue in the United States. We study this phenomenon in this paper by exploring if opinions on climate change alone can produce an accurate model for predicting the 2020 US election. With data collected months before the presidential election, we use three different machine learning methods - K-Nearest Neighbors, Naive Bayes, and Decision Trees - to predict results from the 2020 US presidential and congressional elections. By aggregating our models, we predict the outcome of the presidential election at the state level with 92.2% accuracy. Despite much success, even all of our models were surprised by the Georgia results!

1 Introduction

Climate Change is one of the biggest problems affecting humanity. However, with the political climate in the United States, it has become a political issue. Such division greatly impacts our ability to fight this growing threat. In this report, we show how views on climate change can be used to predict results of the 2020 election.

Before predicting the 2020 election, we first looked at smaller survey data published in 2019 [4]. This data contained state and congressional district wide survey results. Each state and congressional district had a separate data point for Democrats and Republicans, and each data point had 30 parameters- based on responses to a climate change survey. Three classifiers - K-Nearest Neighbors, Naive Bayes, and Decision Trees - were able to perfectly classify the political parties described in the data. This prototyping suggested that opinions on climate change alone may be enough to produce accurate predictions for the 2020 election.

Inspired by success in predicting party affiliation from the 2019 data, we requested 2020 data from [1]. Unlike the 2019 data, opinions on climate change were not separated by party for the 2020 data. For each state, congressional district, and county, we had 20 parameters about the opinions on climate change for the constituents of that area. To predict the 2020 election, we used four classifiers: K-Nearest Neighbors, Naive Bayes, Decision Trees, and Majority Vote. To analyze our results, we used Boxer, an interactive visualization tool specifically designed to compare multiple classifiers [3].

2 Description of the Main Problem

In this project, we predict presidential and congressional results from the US 2020 election on November 3rd solely based on public opinion on climate change. The data was collected through Spring 2020 and published on September 2nd 2020. See [1] for detailed methodology. We provide predictions at both the state and county level for the 2020 presidential election. We also make predictions for the 2020 congressional House races.

3 Related Work

While the correlation between opinions on climate change and voting behavior has been well studied by political scientists in the past, we were unable to find literature that uses machine learning to predict election results based on beliefs on climate change. Our project was inspired by the Climate Change in the American Mind (CCAM) project led by the Yale Program on Climate Change Communication and the George Mason Center for Climate Change Communication [1]. The website (https://climatecommunication.yale.edu/) provides many interesting visualizations of the data from this year and previous years, as well as links to many papers and studies which also used this data. We obtained the data for this project by requesting it via the website. We were unable to, however, find papers that appear to use any form of unsupervised or supervised learning.

4 Dataset

The feature vectors, x, are vectors containing information about the predicted percentage of people in a particular state, congressional district, or county who hold a particular attitude towards climate change. We used data compiled by the CCAM project from Yale [1]. The categories were estimated percentage of those who are and are not interested in:

- whether global warming is or is not happening;
- the causes of global warming;
- the impacts of global warming on their local community, elsewhere in the United States, and around the world;
- actions that are being taken in their local community, the U.S. government, foreign governments, businesses, and presidential candidates in response to global warming

The estimated percentages are based on surveys conducted by the Center over a decade, but we only used the 2020 data.

The labels, or y values, are either Democrat or Republican, based on the results of the presidential (at the state and county level) or congressional (at the congressional district level) races. To obtain the 2020 presidential results by state and county, we used election data from the New York Times election results, [6] which gives the vote (either Democratic or Republican) at both the state and county levels. Washington, D.C. was listed as a state at the state-wide level, so there are 51 data points. For the congressional results, we looked up the incumbents from a Wikipedia page [7] listing all the current representatives and their parties.

5 Four Models for the Main Problem

5.1 Preprocessing the data

The election data obtained from the New York Times and Wikipedia required minimal preprocessing. Election data for all 8 wards in Washington, D.C. were available, but they were combined into one data point and the margin of victory averaged. Most of the preprocessing happened with the data obtained from Yale. We discarded the population data since we wanted to focus on the effect of attitudes towards climate change on voting behavior. For the congressional level analysis, one Libertarian congressional district and five districts with open seats were also discarded. For the county level analysis, the election data on the county level for Alaska was unavailable, so Alaskan data was discarded. Three other counties in Hawaii, Illinois, and Mississippi had no election data, so they were discarded as well.

5.2 Model 1: Distance weighted K-Nearest Neighbors

We choose the Euclidean metric for K-Nearest Neighbors since the variables are continuous. For our choice of K, we use a popular choice, which is the square root of the size of the dataset. For example, for the state data, we choose K = 7 since there are 50 and for the congressional districts data, we choose K = 21 since there are 435 congressional districts. There is no normalization necessary since all the variable values are between 0 and 100. We use distance-weighted K-Nearest Neighbors, namely: $\hat{y} = \frac{\sum_{k=1}^{K} w_{(k)} y_k}{\sum_{k=1}^{K} w_{(k)}}$. Here, y_k is a kth nearest neighbor, $w_{(k)}$ is its weight, calculated by the function $\frac{1}{\delta(x, y_k)}$, where x is the feature vector we are trying to classify and δ denotes the Euclidean distance function. Since $\hat{y}$ yields a float, we convert it into 0 if the result is less than or equal to 0.5 and 1 if the result is greater than 0.5. While distance weighted nearest neighbors, when compared with classical nearest neighbors, does not drastically affect the prediction for states that traditionally always vote for one party, it could be helpful for swing states such as Wisconsin, which had four Democratic nearest neighbors and three Republican neighbors.

5.3 Model 2: Naive Bayes classifier

Since the variables are continuous, for the naive Bayes classifier we modeled each feature as a Gaussian random variable. First, we separate the data into the Democratic samples and the Republican samples. For the j^{th} feature, we compute the MLE mean and variance of that feature for the Democratic samples and the Republican samples, call these $(\mu^j_{\text{Dem}}, \sigma^{2,j}_{\text{Dem}})$ and $(\mu^j_{\text{Rep}}, \sigma^{2,j}_{\text{Rep}})$. Then, to classify a new instance $x_{\text{new}} = [x^1_{\text{new}}, \cdots, x^D_{\text{new}}]^T$ we compute the two naive posteriors

$$\mathbb{P}(\text{Dem}) \prod_{j=1}^{D} \mathbb{P}(x^j_{\text{new}} | \mu^j_{\text{Dem}}, \sigma^{2,j}_{\text{Dem}}), \qquad \mathbb{P}(\text{Rep}) \prod_{j=1}^{D} \mathbb{P}(x^j_{\text{new}} | \mu^j_{\text{Rep}}, \sigma^{2,j}_{\text{Rep}})$$

If the first term is larger, we predict $\widehat{y_{\text{new}}}$ as Democrat. If the second term is larger, we predict $\widehat{y_{\text{new}}}$ as Republican.

One of the underlying assumptions of the Naive Bayes model is that the features are independent random variables. For this data, the features are not independent. A person's view on whether or not climate change is happening will likely be very correlated with their view on whether or not action should be taken by the US government to combat climate change. In spite of this, the Naive Bayes classifier works well in making predictions.

5.4 Model 3: Decision Tree classifier

Statistical analysis of 2019 data revealed that opinions on climate change behave as if they are binary variables. For example, the estimated percentage who think Congress should be doing more or much more to address global warming was below 62% among all Republican opinions while it was above 62% for all Democratic opinions. Therefore, Decision Trees were considered as a possible model for our 2020 election prediction. To construct Decision Trees, all the features were first converted to their nearest integer values. At each node, the algorithm dynamically determined the most informative feature and percentage and used it to split the data. That is, the algorithm considers all features and all possible splits among the data in the node, calculates their information gain, and picks the one with the largest information gain as its next split. The termination condition was when information gain was zero. We then performed 10-cross validation. The ten trees produced for this had an average depth of 24. The overall accuracy was 87%. While pruning did not decrease the accuracy significantly, non-pruned trees were used for the final analysis.

5.5 Model 4: Majority voting classifier

The three previous classifiers used methods that we have covered in class. Inspired by how Random Forests aggregate the decisions of many Decision Trees, we decided to create a fourth classifier that aggregates the results of the other three classifiers. This fourth classifier makes its decision by picking the majority vote of our above 3 models. That is, if at least 2 out of the 3 other models predict Republican (Democratic), the majority model will also predict Republican (Democratic).

5.6 Cross validation

To assess the accuracy of the states' data, the five groups were determined by alphabetizing the states and then splitting them chronologically into five groups of ten. The groups were fairly balanced, so this method of determining groups appears sufficient. The congressional district data, on the other hand, was fairly imbalanced if alphabetized by state and then listed in ascending order by district number. If we split up the list into five equal chunks without reordering the data, we would obtain groups that are mostly one party. Most of the districts in California, the state with the most districts, for example, are Democratic-leaning, so we want to distribute these districts evenly to create more balanced data. To address this issue, we randomly reshuffled the data, and then split into five groups chronologically. We applied the same method to the county data, since if alphabetized by state and then by county, the data exhibited large clusters of Republican-leaning counties (especially in southern states such as Kentucky, which has a surprising number of counties for its size) as well as clusters of blue counties on both coasts. We also randomly reshuffled the counties before splitting them into ten equal groups. We opt for 10-fold cross validation for the county level since there are 3110 counties, so the dataset is much larger than the congressional district and state datasets.

6 Results

6.1 Distance weighted K-Nearest Neighbors

Distance-weighted K-Nearest Neighbors showed that opinions on climate change were fairly accurate in determining voting behavior both in the presidential and congressional races. For the presidential race, five-fold cross validation of the states' data yielded an accuracy of 88.18 percent. Independent runs of 10-fold cross validation on the county level, with random reshuffling, consistently yielded an accuracy of between 91 and 92 percent. The high accuracy at the county level may be attributed to the fact that only 531 of the 3110 counties are Democrat-leaning, so if we were to predict Republican for every county, the accuracy would already be more than 80 percent. At the state level, however, the high accuracy is more significant, since exactly half the states were Democratic-leaning and exactly half Republican-leaning. For the congressional district race, the average of ten runs of 5-fold cross validation, using random reshuffling to create the five equal groups, consistently yielded an accuracy between 87 and 87.5 percent. This accuracy is again significant because roughly half the 435 congressional districts in the house are Democrat-leaning, and the other half Republican-leaning.

6.2 Naive Bayes

On the state level, the Naive Bayes classifier was an accurate predictor of the presidential race in 2020, with 90.2% accuracy on five-fold cross validation data. It also predicted the House races with 88.8% accuracy.

However, the Naive Bayes classifier was not very predictive of the county level results. The classifier achieved 82% accuracy. However, it is worth noting that 83% of counties voted Republican. At the county level, the Naive Bayes predictor is less accurate than a classifier that simply always predicted Republican.

6.3 Decision Tree

The Decision Tree classifier achieved 78.4% accuracy on the state level presidential election, 80.3% accuracy in the congressional house races, and 88.2% accuracy in the county level presidential election results. All ten trees considered the parameter "Estimated percentage who are not interested in actions that the presidential candidates plan to take in response to global warming" to be most informative with the split percentage of 23%.

6.4 Majority Vote

The majority vote classifier was the most accurate predictor in every race. It predicted the presidential race at the state level with accuracy 92.2%, the congressional house races with accuracy 89.7%, and the county level presidential election results with accuracy 91.2%.

6.5 Summary of accuracy of each classifier

In the table below, we summarize the accuracy of each classifier using cross validation data for the 2020 election. In all three races, the majority vote classifier yields the highest accuracy.

Cross Validation Accuracy				
	Naive Bayes	K Nearest Neighbors	Decision Tree	Majority Vote
States, Presidential	90.2%	88.2%	78.4%	92.2%
Congressional Districts, House	88.8%	87.4%	80.3%	89.7%
Counties, Presidential	82.0%	91.2%	88.2%	91.2%
Overall Accuracy	87.0%	88.9%	82.3%	91.0%

6.6 Summary of pairwise classifier consensus

While the overall accuracy of the classifiers were similar to each other, there were some discrepancies in the predictions for each instance. We summarize this in Figure 2, which was produced using the classifier visualization tool Boxer [3]. We also analyzed confusion matrices produced by Boxer, which give a more detailed breakdown of the false positive and false negative rate. These matrices reveal that Decision Tree and Majority Vote work especially well for predicting Republican (the majority party on the county level) votes correctly, while K-Nearest Neighbors and Naive Bayes showed more balanced performance regardless of the political parties.

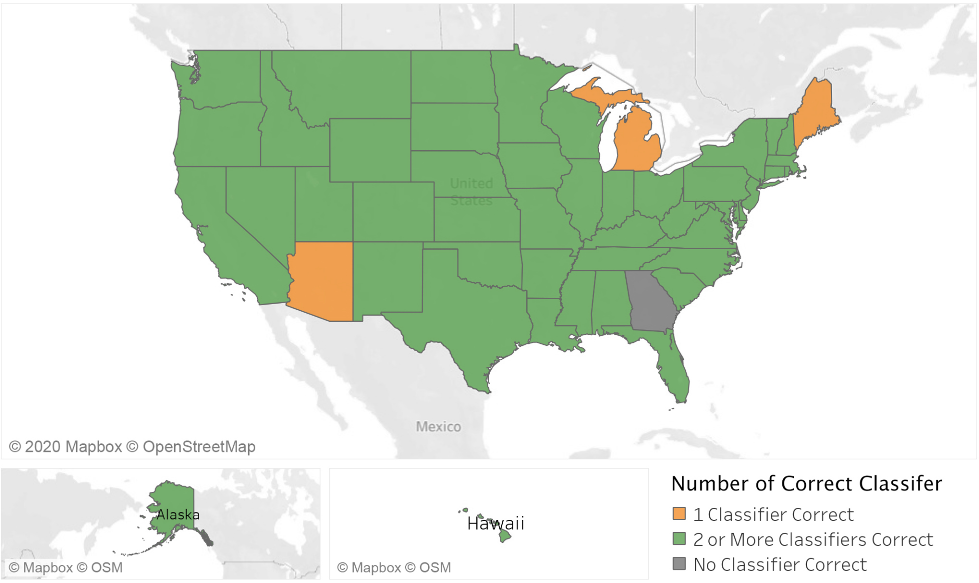

Figure 1: Of three main classifiers - Naive Bayes, K-Nearest Neighbors and Decision Tree - Georgia was the only state that none of the classifiers predicted the election correctly. Three states - Arizona, Maine, and Michigan - only one of the three classifiers were correct. For the remaining 47 States (including District of Columbia), two or three classifiers were correct. This means the majority vote classifiers is 47/51=92.2% accurate

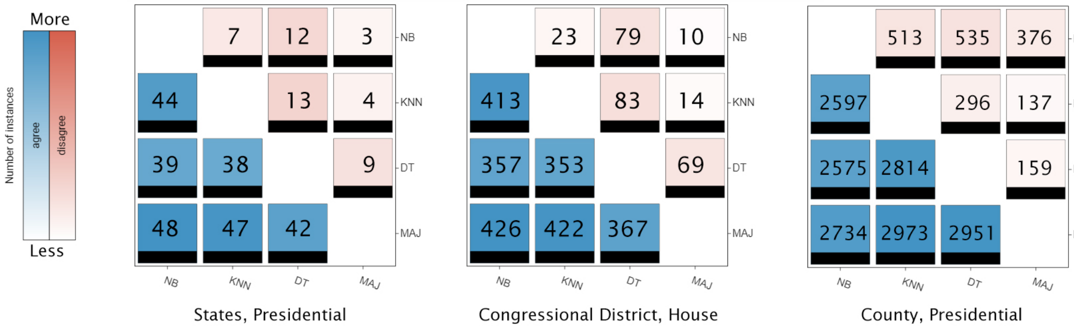

Figure 2: The chart shows pairwise classifier consensus. The vertical horizontal axes shows the names of the classifiers: NB = Naive Bayes, KNN = K-Nearest Neighbors, DT = Decision Tree, and MAJ = Majority vote. The red boxes show the number of instances that two classifiers disagree, and the blue boxes show the number of instances that two classifiers agree.

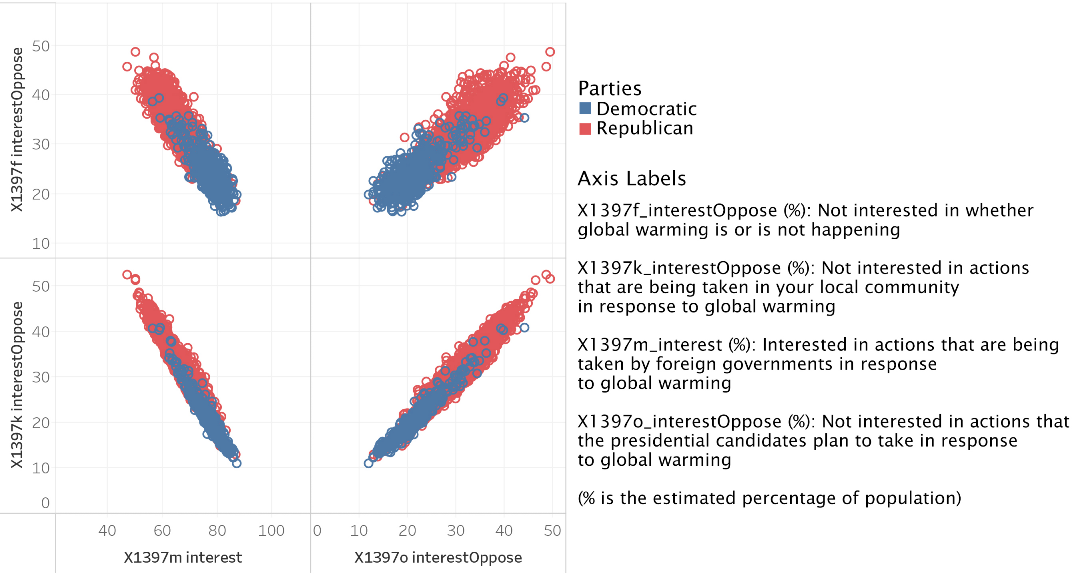

Figure 3: Results from a sample of the 2020 survey questions. The high degree of clustering between the Democratic responses and the Republican responses could be a significant reason for the success of the K-Nearest Neighbors classifier.

7 Conclusions

There are many conclusions we can draw from these classifiers. First and foremost, climate change is an incredibly partisan issue in the US. By solely looking at voters' views on climate change, we were able to predict many electoral races with over 90% accuracy! A different 2018 survey in [2] reveals that Republican opinions on climate change have changed little since 2013 while Democrats show an increase in worry.

Figure 1 shows that Georgia is the only state that all classifiers incorrectly predicted the 2020 presidential state level election result. Georgia is also the state that many comprehensive predictive models from major news outlets missed, including the one employed by the Washington Post [5].

It is worth noting that the Majority Vote classifier was the most accurate classifier, achieving the highest accuracy for all three sets of election data. This is, in part, because not every classifier is perfect; where one particular classifier might be incorrect, the other two classifiers often "pick up the slack" and can correct the mistakes of the third classifier. For example, in our classifier analysis using the visualization system Boxer, we observed that the Decision Tree and K-Nearest Neighbors caught many misclassified instances of the least accurate classifier at the county level. Therefore, the Majority Vote classifier was still able to achieve the same accuracy as the K-Nearest Neighbors classifier.

In our accuracy analysis, we also noticed that K-Nearest Neighbors was able to achieve the highest average accuracy (aside from the Majority Vote classifier). Our statistical analysis using Tableau reveals that when we compare two parameters at a time, the data tends to form two clusters- one for each party (see Figure 3). This may account for K-Nearest Neighbors being the most reliable classifier for our data. For the congressional race, K-Nearest Neighbors and Naive Bayes both yielded about 89% accuracy while the Decision Tree was the least accurate, yielding about an 80 % accuracy. Nonetheless, the Decision Tree caught some of misclassified instances of two more accurate classifiers, so overall, Majority Vote was the best classifier in this case, with an accuracy of about 90%.

At the county level, the Naive Bayes classifier was the least accurate. Perhaps this is due to skewed data. As 83% of counties voted Republican, this classifier was less accurate than the predictor which always guessed Republican. Looking more deeply into this data, the Naive Bayes classifier is the most likely to be correct when it makes a Democratic prediction, with over 88.7% accuracy on these counties, compared to 67.4%, 70.4%, and 80.4% for K Nearest Neighbors, Decision Tree, and Majority respectively. The formula used to make predictions for the Naive Bayes classifier takes into account the small probability for a Democratic county (compared to the large probability for a Republican county) and is thus best able to make predictions for the small percentage of Democratic counties. One way to address some of the inaccuracies of the Naive Bayes classifier would be to use a more complicated Bayesian Network, taking into account the correlation between the different features.

8 Future Work

In the future, we could go a step further and try to use voter attitudes towards certain issues to predict margin of victory, which indicates how competitive an election was. Margins of victory, floats between -100 and 100, are also given by the New York Times county data. In our dataset, the margin of victory is calculated by subtracting the percentage of the vote Biden received from the percentage of the vote Trump received. For example, in Dane county, Biden received 75.5 percent of the popular vote while Trump received 22.8 percent, which yields a margin of victory of $22.8 - 75.5 = -52.7$.

We tried to use local linear regression to predict the margin of victory at the county level using the Yale climate data. A natural choice for calculating margin of victory is local linear regression, which is inexpensive to run and designed to predict float values. We can use the 56 of the nearest neighbors to construct a feature matrix $\mathbf{X}$, along with a vector of labels $\mathbf{y}$. Given a county's feature vector $\mathbf{x}$, we use the formula given in class, $\hat{y} = \mathbf{x}^{\mathbf{T}}\hat{\theta}$, where $\hat{\theta} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$. In order to assess accuracy, we can calculate the absolute value of the difference between the actual margin of victory and $\hat{y}$ and take the averages of these values.

The climate data was not very effective at predicting county voter margins, as to be expected since voters consider a wide range of other issues while voting. On average, the local linear regression prediction was off by 37.92 percentage points. The prediction was slightly more accurate for the Democratic race, at 21.84 percentage points, compared to an average error of 41.23 points for the Republican race. The best prediction was an error of 9.688×10^{-3} in Tippah Mississippi, one of many Republican-leaning districts in the state where the citizens are not concerned with the effects of global warming. The worst prediction was an error margin of 101.58462134 in the Republican leaning Williams, Ohio, where the algorithm also predicted the wrong party. In Williams, citizens are very concerned about global warming; 69.665 percent are worried about it, but clearly, climate change was not their number one concern when casting ballots.

In light of these results, a more accurate prediction of margin of victory clearly requires a tweak in methodology, or gathering further population and opinion data. There is much more future work that could be done if we had similar data involving a wide range of issues - attitudes about COVID, the economy, foreign policy, race relations, etc. Such information would likely help in predicting margin of victory. But also, further analysis could be performed to analyze which issues were important to different groups of voters across the country. Democrats might not be able to persuade the Republican voters of Williams County, Ohio with talk of climate change. But knowing which issues accurately predict votes in similar areas would be invaluable for a Democrat running in Williams County.

9 Appendix

Source code (GitHub Repository): https://github.com/yandiwu/CS760project

References

[1] Matthew Ballew, Jennifer Marlon, and Anthony Leiserowitz. *Explore Climate Change in the American Mind.* URL: https://climatecommunication.yale.edu/visualizations-data/americans-climate-views/?fbclid=IwAR2uHbaCDvfYQk3A6Zw0yKHCwoxaYxH3YGqm0ONY_Tp_wYCxAT_lUZ5_Azg. (accessed: 11.15.2016).

[2] Moira Fagan and Christine Huang. *A look at how people around the world view climate change.* URL: https://www.pewresearch.org/fact-tank/2019/04/18/a-look-at-how-people-around-the-world-view-climate-change/.

[3] Michael Gleicher et al. *Boxer: Interactive Comparison of Classifier Results.* URL: https://graphics.cs.wisc.edu/Papers/2020/GBYH20/.

[4] Matto Mildenberger et al. *Democratic & Republican Climate Opinion Maps 2018.* URL: https://climatecommunication.yale.edu/visualizations-data/partisan-maps-2018-old/?est=happening&group=dem&type=value&geo=cd. (accessed: 11.15.2016).

[5] The Washington Post. *We predicted the states Biden would win 100 days before the election.* URL: https://www.washingtonpost.com/politics/2020/11/12/we-predicted-states-biden-would-win-100-days-before-election/.

[6] The New York Times. *Presidential Election Results: Biden Wins.* URL: https://www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html. (accessed: 11.18.2020).

[7] Wikipedia. *2020 United States House of Representatives elections.* URL: https://en.wikipedia.org/wiki/2020_United_States_House_of_Representatives_elections. (accessed: 11.18.2020).