STAT 605 Group Project - Final Draft

Hanlin Tang(htang79), Hua tong(htong24) , Yuechuan Chen(ychen959), Yuhan Zhou(zhou453)

December 10, 2020

1 Introduction

As the recent 2020 election ended, some interesting questions occurred. We decide to study some of those questions with the ANES election data, which contains information and interview response of voters.

The very first question we want to look at is the 'shifting' phenomenon, which means that some people declared to vote for somebody, but eventually voted someone else. We are wondering which kind of characteristics or pattern are related to such 'shifting'. Due to the relative high dimension of the data set, we did our analysis mostly with PCA, and also built a SVM model for prediction.

We also try to predict the vote of each people given the data from the pre-review by linear regression and tree method. We also use ANOVA and T-test to find out some factors that may affect the result.

2 Data

2.1 Source Data Description

We use the ANES election data from the American national election studies (https://electionstudies.org/data-center/anes-time-series-cumulative-data-file/).

The data is about the personal information and political positions of the interviewed people. Each row is a interviewee, and each column is an answer towards an interview question. Also, most questions have a categorical (nominal or ordinal) answer.

The original data has 59944 observations (interviewees), ranging from year 1948 to 2016. In this project, we focus on year 2016. Also, the original data has 1029 columns (variables).

There are two columns that worth noting here:

1. VCF0704 is about who the interviewee voted, 1 for dem and 2 for rep (0 for other cases, including missing or refusing to answer).

2. VCF0734 is the "Intended Presidential Vote versus Actual Presidential Vote". it's value can be summarized with the following table:

Intended vote Actual vote	Dem	Not Decided or Others	Rep
Dem	1	2	3
Not Decided or Others	4	5	6
Rep	7	8	9

These two columns are the y(s) we are going to focus on. And here is a summary of these two variables: (among the cleaned data, see below)

1. VCF0704: 1349 for 1(dem), 1238 for 2 (rep), and 1061 NAs.

2. VCF0734:

values	1	2	3	4	5	6	7	8	9	NA
counts	1130	80	14	168	201	119	27	70	999	840

2.2 Data Clean

First of all, we only keep data of year 2016, which has 4270 rows. Then, we keep data with interview response result (there is a variable that indicate whether this interviewee has response result, we simply keep data with that field to be true), which has 3648.

Then, since the data is about a long time range and the questions changed a lot of times, which leads to many empty columns or meaningless columns, we drop all columns that is all NA or all the same value from the data above and reached a shape of 3648*341.

We also drop those columns that is identical with other columns (for example, there are two columns: pre-interview and post-interview response, which is the same for 2016, so we only keep one. We do this by check the correlation. Most columns are numeric so this worked well. We checked those that does not support correlation calculation manually.) Then we get a shape of 3648*337.

In our next step, we first replace some 'value NA' to 'actual NA': for example, in some columns, 9 refers to NA. After that, due to the long scope of this data source, there are many columns that is actually covering same information and redundant. For example, VCF0706 refers to the actual voting result, with 5 levels, while the VCF0705 also refers to voting result, with only 3 levels and can be derived from VCF0706. We manually drop them by checking the definition.

2.3 Imputation

For the original data, as we mentioned before, there are 337 variables. Among them, 254 variables have NAs. We collect the number of NAs among different variables, and find that the median number of NAs is 32 and the 3rd quartile of the distribution of NAs is 85. However, the mean is 272.2 and the maximum value is 3364. Therefore, the pattern of missing values is highly skewed, so we need pre-selection before the imputation process.

We drop ten variables with more than 1824 NAs, which means more than half of samples have NAs in these variables. We believe that this would cause inaccuracy in the imputation process, so we would simply remove them. Later, we remove 32 meaningless traits, like weights and time spent to decide who to vote. We do this to simplify the imputation process. So far, we only keep the individuals with valid value in our target variables VCF0704 (who did you vote) and VCF0734 (comparisons between actual voting and intended voting). After the pre-selection, we do the imputation with two methods: Predictive Mean Matching and Random Forest Imputation.

2.3.1 Predictive Mean Matching

The predictive mean matching algorithm selects the closest observed values (typically three cases) to the missing value . Since the structure of the data is very complex, both the numeric and categorical variables have missing values, and the method can work on both kind of the variables, we choose it as our candidate method.

We implement this method with the function mice() in R. Considering the huge amount of computation, we use the computing server to do the imputation. We allocate the virtual memory of 794 MB and resident set size of 378 MB. The calculation takes about 1.5 hours to complete.

However, we discover that for a three level variable which is about the attitude towards government, there are always NAs which are not able to be imputed. Due to this failure, we plan to try the random forest method.

2.3.2 Random Forest Imputation

The random forest imputation uses the random forest algorithm to predict the NAs while using the observed data as the training set. Since it works for the numeric and discrete variables and is well-known for its predicting power, we decide to use it to impute the missing values.

We implement this method with function missForest() in R, the maximum number of iteration was set to 10. We also use the computing server to do the calculation. We allocate the virtual memory of 710 MB and resident set size of 295 MB. The calculation takes about 1 hour to complete.

The unsolved NAs in the PMM method are imputed by random forest method, so we decide to use the results from random forest imputation as our input data for analysis.

3 Method

In this problem, we propose several methods for different perspective:

3.1 Principal components analysis (PCA)

As we know, Principal components analysis (PCA) can do dimension reduction by using a new basis in a smaller dimension, where axis are those directions with the most variation in original X space. Here, for better visualization, we mostly keep 2 dimensions left, and we will soon see that 2 dimensions are sufficient. The ellipse is the confidence interval on both dimension, it's generated to provide a clear view of the grouping status.

3.2 SVM

Later, we built a model for prediction with support vector machine (SVM). The support vector machine is a widely used supervised learning method and was used to do the prediction for political position switch (intended presidential vote contradicts with actual presidential vote) in our project, since there are 8 levels for the political position switch, we believe SVM is the best fit methods for this problem.

3.3 Linear & Tree Regression

As for the prediction of vote of each people, we decide to use linear regression and tree method. We select 35 important variables for this part, and split train and test data. Then we use the 'step' function in R to do linear model selection and the 'rpart' package to fit a tree model. For model evaluation we display the confusion matrix to judge the accuracy. Finally, we do ANOVA and T-tests to look for some other key factors.

4 Findings

4.1 The myth of "The silent majority"

It's a generally known a fact that most of the polls and media performed extremely poor while predicting the results of 2016 presidential election. Someone proposed an idea that, such failure had two reasons:

1. People tend to lie about their support for Donald Trump;

2. People without college degree, especially white people without college degree, were not fully covered in the poll.

We planned to analyze both points above, however after data cleaning, quality control and the imputation process, there's only 400 points left. The data points for the analysis of the first point is to small to do . So, we decides to only analyze the second point in our current project.

Political position and voting preference is strongly correlated with individuals non-political traits like education, income and socioeconomic status. So we would like to constructs PCA on the non-political to see if we can distinguish people of different political preference with this. In the plot (Figure 1), blue stands for people who voted for Democratic candidate and red stands for people who voted for Republican candidate. Clearly, there's no obvious differences between the two group, therefore, we decides to look into the DEM and GOP supporters separately.

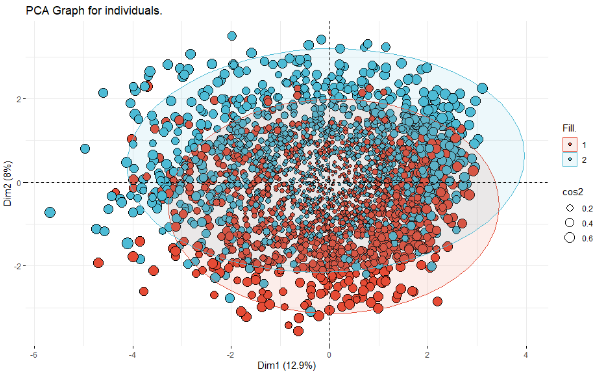
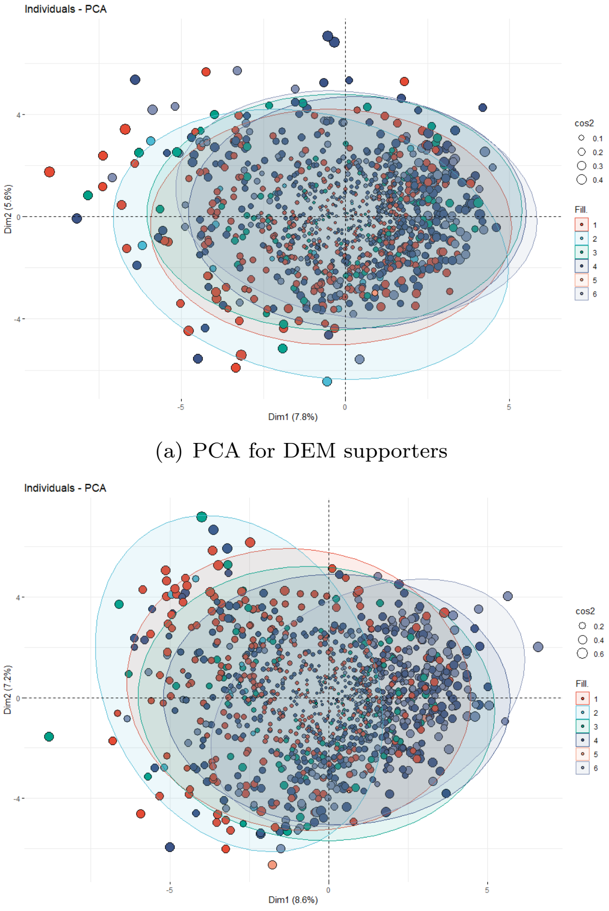


Figure 1: PCA based on non-political traits



(a) PCA for DEM supporters

(b) PCA for GOP supporters

Figure 2: PCA for separate groups

By conducting PCA analysis on the both sets (Figure 2) we discovered following patterns:

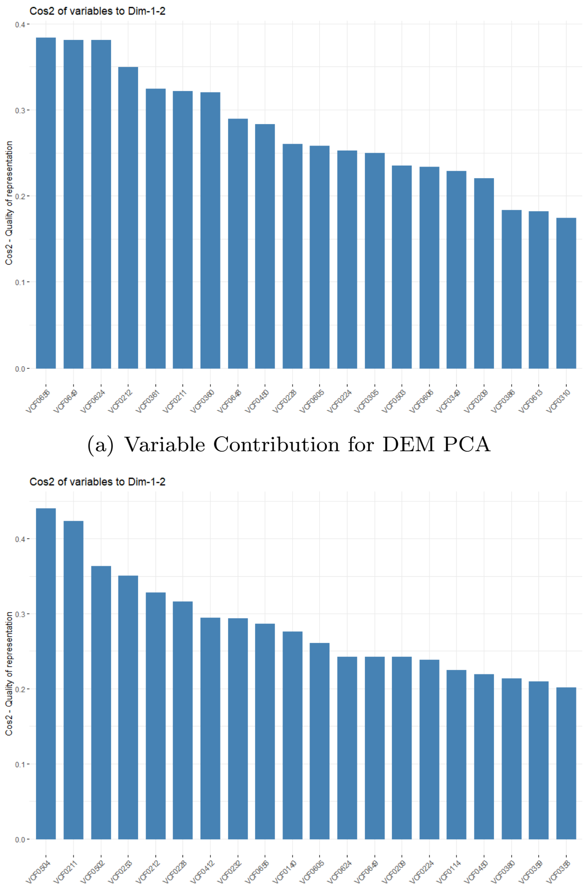
1. For the DEM supporters, the distribution of people generally follows our common sense, while the distribution of GOP supporters is strange;

2. The distribution of variables' contribution to the PCA analysis varies a lot between the DEM and GOP supporters.

Firstly, let's examine the grouping results of PCA. The number of the label (that is the different color in Figure 2) stands for:

1. Average working class;
2. Neither average or upper working class;
3. Upper working class;
4. Average middle class;
5. Neither average or upper middle class;
6. Upper middle class;

While for the DEM PCA, it shows that almost all class are grouped together and sharing the same center, which clearly don't follow our understanding about the logic behind the voting preference in each class. So, as we can see from the ellipse's x axis and y axis, though the pattern is not so obvious, for the GOP supporters, the upper and lower middle class are grouped together and the working class only share some similarities with the middle of the society, which doesn't contradicts with our common sense.



(a) Variable Contribution for DEM PCA

(b) Variable Contribution for GOP PCA

Figure 3: Variable Contribution for separate PCA

Secondly, for the contribution of the variables, the top 3 variables for the DEM PCA is:

1. Trust in Government(VCF0656);

2. Government Responsiveness(VCF0649);
3. How Much Elections Make Government Pay Attention to People(VCF0624)

And the top 3 variables for the GOP PCA is:

1. Republican Party- Guaranteed Jobs-Living Scale(VCF0514);
2. Feelings to liberals(VCF0211);
3. Which Party Favors Guaranteed Jobs and Standard of Living (VCF0512);

It's obvious that the variables that decide the PCA are completely different. For DEM, the variables are more sociological, which focus on the government. While the variables for the GOP are much more about, which party can guarantee jobs, which is a well-know varibale that's related to socioeconomic status.

Such huge differences can be viewed as a proof for the "silent majority" myth. For, if there are no surveyed bias in population group or class, then the contribution of the variables won't change this much, since the surveyed class or groups between each political groups(as the hypothesis stated) and the logic behind voting, "choosing the candidates who can serve the interest of me", don't change in different group, the only difference would be on the distribution of the value for each variables, not the contribution.

Though further examination should be made, like conducting more rigorous statistical test on this, we would still say it's a really interesting discovery.

4.2 Prediction of Political Position Switch

ANES has a very interesting variable which called "Intended Presidential Vote versus Actual Presidential Vote", which contains the information about the individuals intended voting and actual voting. We decide to make a prediction tool that can help us identify these voters who tend to change their position. These "Swing voters" should be viewed as the key group and need to be taken carefully just like those swing states.

We used Support Vector Machine for the analysis, the Y is the VCF0734 we mentioned above, without level 5 (that is, 8 levels left).

4.3 Position Switch Inference

(In this section, the label and color are from the variable VCF0734, which you can find the definition in section 2.1)

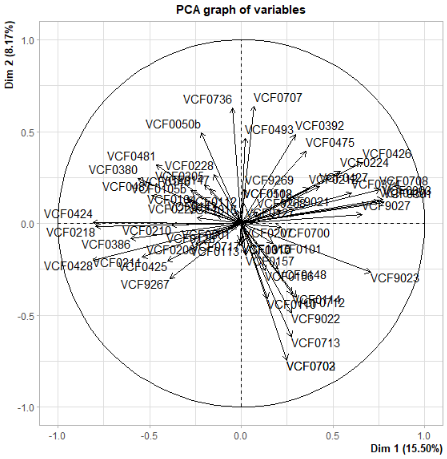
From picture a in fig 5, we can see that those who 'INTENDED' to vote for rep/dem and did that (group 1 and 9) are on the two side of the dimension 2. And yet those who 'INTENDED' to vote and did not vote (group 4 and 6) are tend to have higher value on dimension 1, and we can see the dem suffered more from this group. Those who are at the middle tend to fall into the group who didn't decided or who changed their mind.

After taking away those 'loyalty voters' and take a closer look about those 'middle voters', we can get the fig 6. From it we can see that, still dimension 2 can determine the political position. Meanwhile, higher value in dimension 1 would result in more likely of 'silence', that the voter eventually did not vote or voted 'other'.

Moreover, there are some interesting things about these voters:

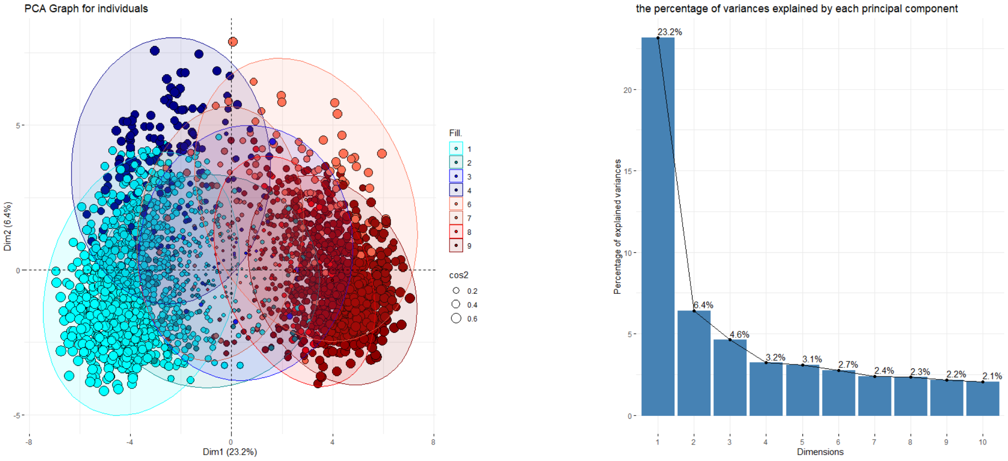
1. There are large amount of voters who decided to "abandon" to vote(vote other parties or quit voting) for both of the party, and by viewing the effect of each variables, the knowledge of politics(VCF0050b) have big influence on this, this phenomenon might corresponds to the conspiracy theory and incidents before the 2016 election, people who have little knowledge of politics tended to believe conspiracy and would quit voting if they thought those conspiracy theory or follow the media's report about the mail gate;

2. As always, the middle voters should be taken into account, as many middle voters "betrayed" their intended party or joined the election eventually, this phenomenon might be much more interesting in the 2020 election.



(a) Radar figure for PCA

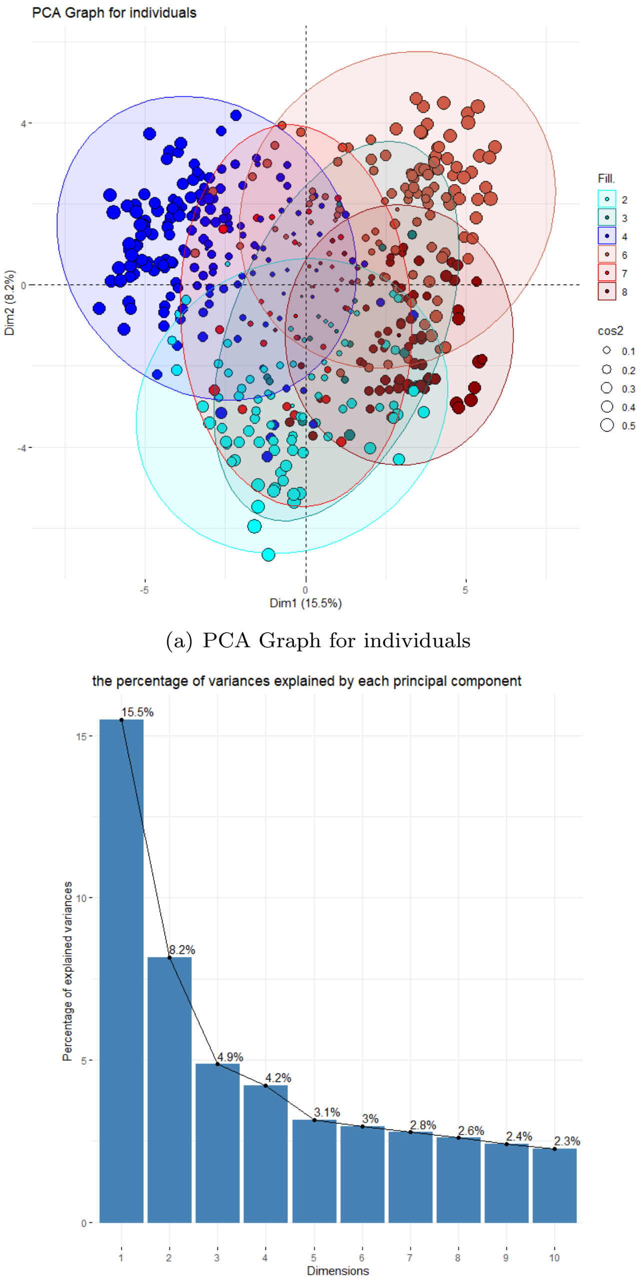
Figure 4: PCA2



(a) PCA Graph for individuals

(b) the percentage of variances explained by each principal component

Figure 5: PCA1



(a) PCA Graph for individuals

(b) the percentage of variances explained by each principal component

Figure 6: PCA2

4.4 Prediction of votes

We extract the people that vote for Democratic or Republican, and ignore the others who vote for other candidates or did not vote. The variables we select include age, number of children, income level, thermometers for different people and parties, party identification, etc. Before fitting the model, we split train and test data in order to evaluate our model performance.

Since the pre-interview include a variable of party identification, we guess it would not be difficult for us to predict the votes, and the result of linear regression verifies our point. Some variables related to the attitude towards parties and candidates appear to be significant factors. We also find some interesting variables, like the attitude to liberals, Muslims and Christians, and whether parents are native born, are also key factors. The accuracy of our prediction is about 92%, which means that votes of most people can be predicted from the pre-review. The confusion matrix is displayed below:

Confusion Matrix: linear regression		
	vote: Democrat	vote: Republican
predict: Demo	298	20
predict: Repub	23	268

Then we try tree method to do model fitting. The plot of tree is shown below together with the confusion matrix. Here are some explanations of the split points:

1. VCF0301: Party identification range from 1 (Democrat) to 7 (Republican);
2. VCF0374: Whether likes anything about the Democratic party;
3. VCF0386: Whether likes anything about the Republican party;
4. VCF0218: Thermometer for Democratic party.

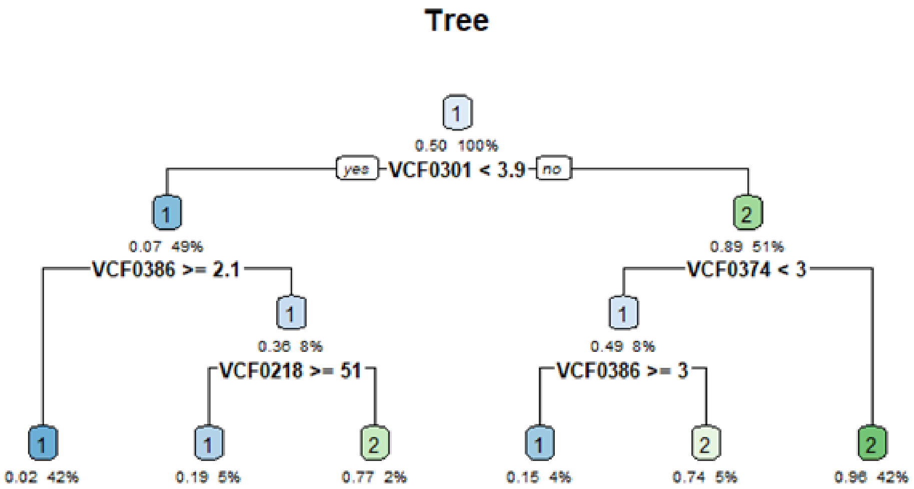


Figure 7: Tree plot

Confusion Matrix: tree method		
	vote: Democrat	vote: Republican
predict: Demo	291	21
predict: Repub	30	267

Then we do ANOVA and T-tests to search for some other factors related to the votes. The result shows that:

1. People with different levels of political information make different decisions;
2. Elder people are more likely to vote for the Republican party;
3. Male people like to vote for the Republican party, while females like to vote for the Democratic party;
4. People with different levels of education have different votes;
5. People in different regions make different votes;
6. The work status could affect the vote;
7. People of different religions vote differently: Protestants like to vote for the Republican, while Jewish like to vote for the Democrat;
8. Votes of people of different marital status are slightly different;

9. People are very likely to make the same decisions as previous elections.

However, our previous prediction seems a little unpersuasive since the pre-review contains too much party information, which is a strong signal of the vote. Now we try to predict the votes by only basic information of people, including gender, education, religion, marital status, etc. To do this, we remove all variables that are related to the parties.

First we try the linear regression method. After model selection, some significant variables are church attendance, home ownership, gender and education. The R-square is about 0.11, and the accuracy of prediction is about 63%. We can see that the model still works, but the effect is not that good. The confusion matrix is showed below:

Confusion Matrix: linear regression 2		
	vote: Democrat	vote: Republican
predict: Demo	224	127
predict: Repub	97	161

Then we apply the tree method. One advantage here is that tree method could handle non-ordinal variables, which is not appropriate for linear regression. The tree plot is displayed and the split points are:

1. VCF0105a: Race and ethnicity;
2. VCF0128: Personal religion;
3. VCF0130: Church attendance;
4. VCF0110: Education level.

By the variable importance table, we find that race (ethnicity) and religion are the top 2 influential variables. Since the tree method could make use of the information, it is obvious that it performs better than linear regression. The prediction accuracy here is about 71%.

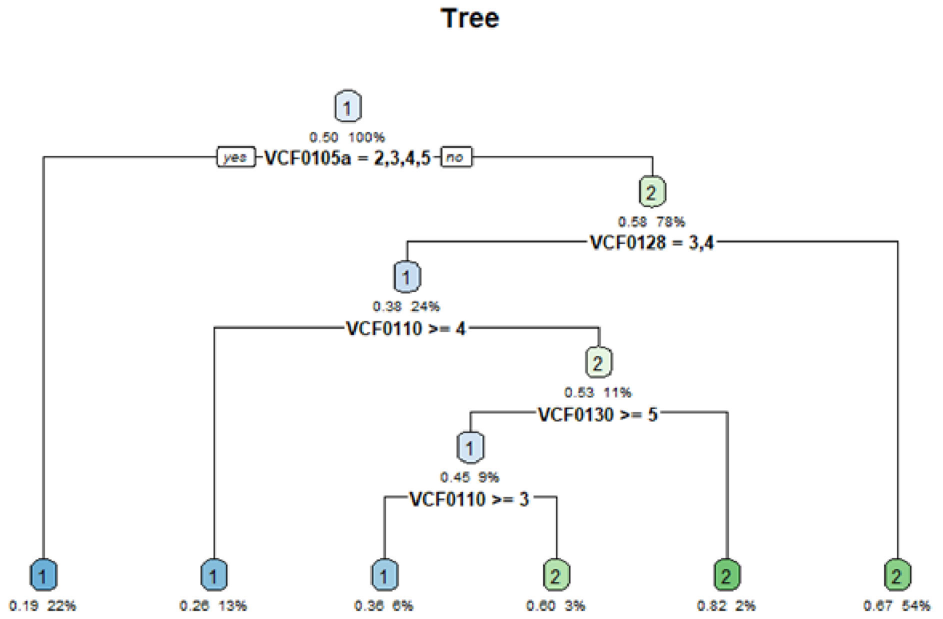


Figure 8: Tree plot 2

Confusion Matrix: tree method 2

	vote: Democrat	vote: Republican
predict: Demo	193	50
predict: Repub	128	238

Then we decide to apply tree method to some previous data to evaluate its performance. We extract such variables from the data of 2012, 2008, ... , 1960, and then fit the model to find important variables and prediction accuracy. According to the results, the model makes 69% and 71% accuracy in the elections in 2012 and 2008. But this number fluctuates around 65% in earlier years, which indicates that the prediction is easier in recent elections. As for the important variables, VCF0105a (race and ethnicity) is always the top 1. It explains most of the variability in votes.

For this part, first we run 2 linear regression jobs, which takes about 1 minute each. Then we use the CHTC to run 14 parallel jobs, which take about 8 minutes in average, to do prediction for the 14 previous elections.

5 Summary and Conclusions

To summarize this, we proposed two questions: to analysis the position shifting, and to predict based on pre-election information.

For the first problem we maninly used PCA, and combined with SVM to analysis. For the second problem, we used linear tree regression as our model.

For the analysis problem, we found how the different groups's distribution in voting, and in shifting. Though points like the proof of silent majority may need much more rigorous statistical test, the tendency deserve our focus and we should look into this.

For the prediction problem, we reached an prediction accuracy of 71% and learned the relationships between some variable and the result.