STAT 605 Group Project - First Draft

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

November 30, 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: that is, those who declared to vote for somebody, eventually voted someone else. We are wondering which kind of characteristics or pattern can relate to such 'shifting'.

Due to the extra high dimension of the data set, we did our analysis mostly with PCA, and also had a SVM model for prediction.

Then we 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 Details of methods and data

2.1 Data

-

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 position of the interviewed person.

There original data set has 1841 variables and 4270 observations for 2016. Since the variables are too many, We first manually selected 120 among those variables based on whether they are related to our question. We further selected those data in 2016 and has a post-interview result, and dropped those that has no post-election interview data, no voting outcome data, no social class data, no gender data. We end up with a data set with 2587 rows and 65 columns. (And we drop many columns that has all NA.)

Since there are lots of NAs in each columns, we use random forest to impute the the missing values, this helped us saving the size of the dataset.

In the end we included variables like these: attitudes towards candidates, presidents and vice presidents of the two parties, work, family, religious situation, social class, attitudes to different races, vote in the last election, and the tendency for early elections, etc.

2.2 Method

-

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 is sufficient with screeplot.

In general, the way we use PCA to analysis is to first hold the interested variable, here is whether this subject shifted opinion, as well as some other columns that is either empty or not reasonable to be included, out of our X. And perform PCA on this X. After that, we draw them in the new 2-d space, and then color them with our interested variable.

Later, we also built a model for prediction with support vector machine (SVM). The support vector machine is a widely used supervised learning methods and was used to solve the prediction of 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.

SVM classifies the datasets by constructing hyperplanes in a high dimensional se and try to find a hyperplane that has the largest distance to the nearest training-data point and the mapping to the hyperplanes is usually related to the kernel function, which decides the type of mapping and the type of the classifier, by finding such hyperplanes, we build the corresponding classifier.

As for the prediction of vote of each people, we decide to use linear regression and tree method. We select 35 important variables that are ordinal 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.

3 Finding

3.1 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 planed to analyze both points above, however after data cleaning and quality control, we don't have enough data points for the analysis of the first point. 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, blue stands for people

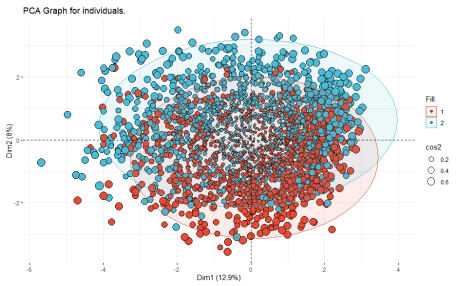
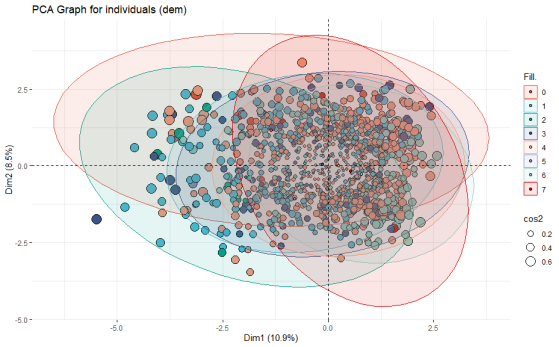
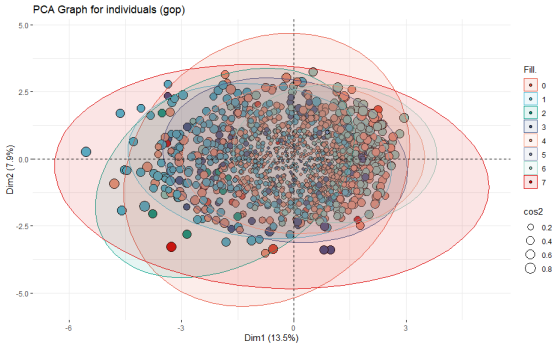


Figure 1: PCA based on non-political traits

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.



(a) PCA for DEM supporters



(b) PCA for GOP supporters

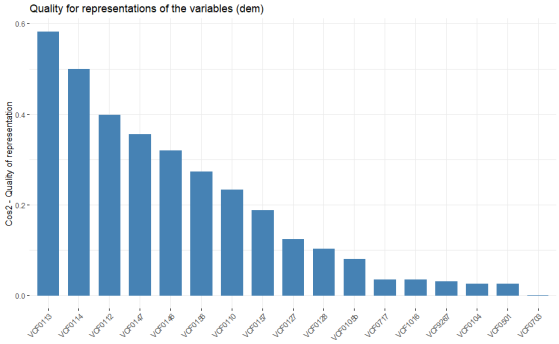
Figure 2: PCA for separate groups

By conducting PCA analysis on the both sets we discovered following patterns:

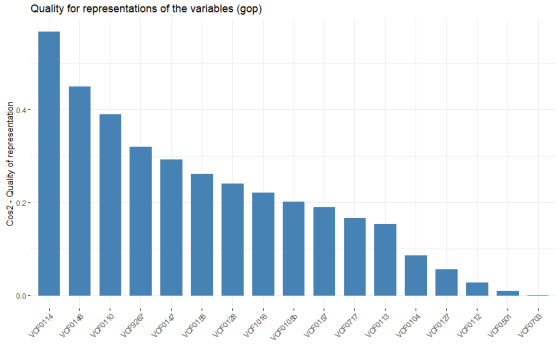
1. For the DEM supporters, the distribution of people generally follows our common sense, while the distribution of GOP supporters is strange;

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.,For the label, larger the number is, higher the socioeconomic status of the individuals. So, as we can see, for the dem supporters, the working class and the lower middle class are grouped together and the upper middle class, high class and the low class only share some similarities with the middle of the society, which doesn't contradicts with our common sense. While for the GOP 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.



(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. Political region:south or north(VCF0113);

2. Income group(VCF0114);

3. Geographical Region:South, North Central, North East, West(VCF0112)

And the top 3 variables for the GOP PCA is:

1. Income group(VCF0114);

2. Home Ownership(VCF0146);

3. Education level(VCF0110);

It's obvious that the variables that decide the PCA are completely different. For DEM, the variables are more sociological, which focus on the environment. While the variables for the GOP are much more about the economical status and educational level, which is a well-know varibale that's related to socioeconomic status. Further, by examining the variable contribution for GOP PCA, variables like Political Region and Geographical Region are very insignificant. 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.

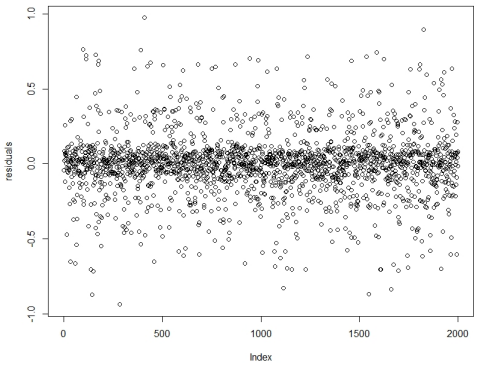
Though further examination should be made, like conducting more rigorous statistical test on this, we would still say it's a really interesting discovery.

3.2 Prediction of votes

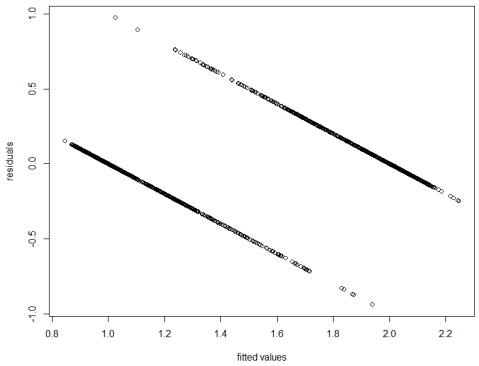
We extract the people that vote for Democratic or Republican, and ignore the others who vote for other candidates or did not vote. The variables we select include age, number of children, income level, thermometers for different people and parties, party identification, etc.

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 and the plots of residuals are displayed below (1 stands for Democrat and 2 stands for Republican in the data):

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



(a) Residuals

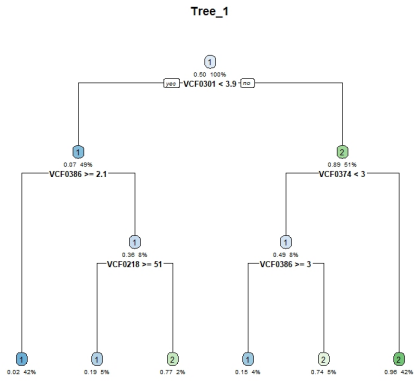


(b) Residuals v.s Fitted values

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.

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



(c) Tree plot

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 female people 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.

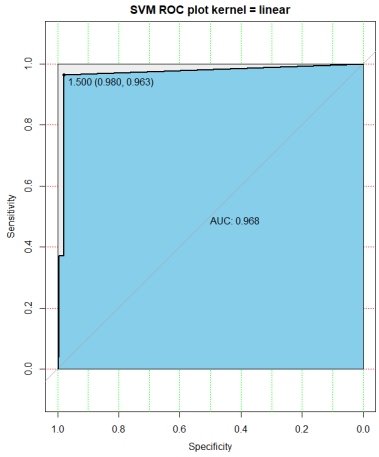
3.3 Prediction of Political Position Switch

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.

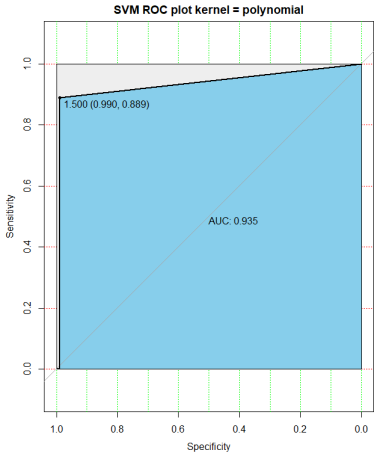
We used Support Vector Machine for the analysis, for we have 8 levels to classify. The levels are:

1. INTENDED Democratic: voted Democratic
2. INTENDED undecided: voted Democratic; INTENDED 'other' party: VOTED Democratic
3. INTENDED Republican: voted Democratic
4. INTENDED Democratic: did not vote/voted 'other' party

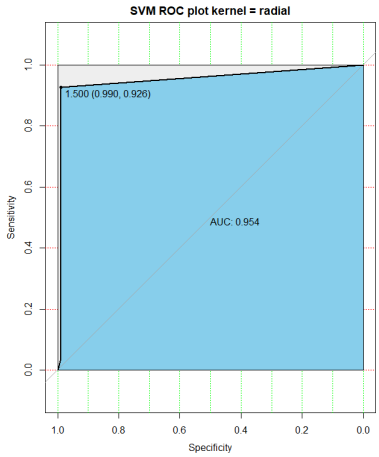
5. INTENDED Republican: did not vote/voted 'other' party
6. INTENDED Democratic: voted Republican
7. INTENDED undecided: voted Republican; INTENDED 'other' party: voted Republican
8. INTENDED Republican: voted Republican



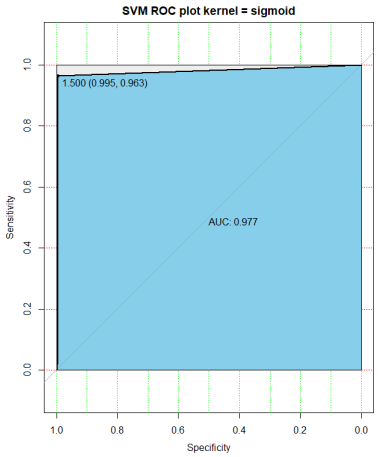
(d) kernel=linear



(e) kernel=polynomial



(f) kernel=radial



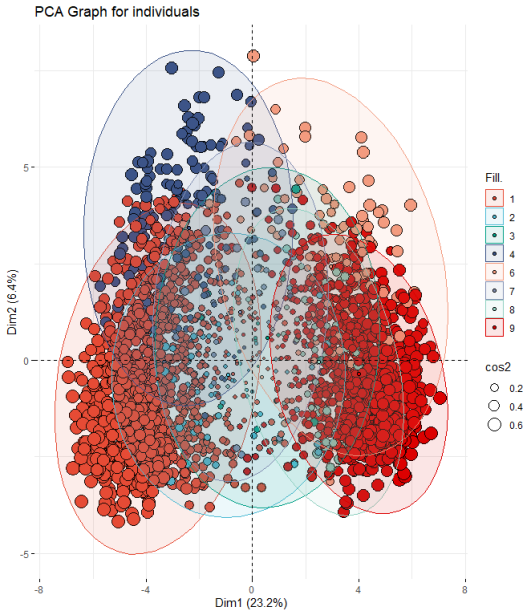
(g) kernel=sigmoid

Figure 4: SVM with Different kernels

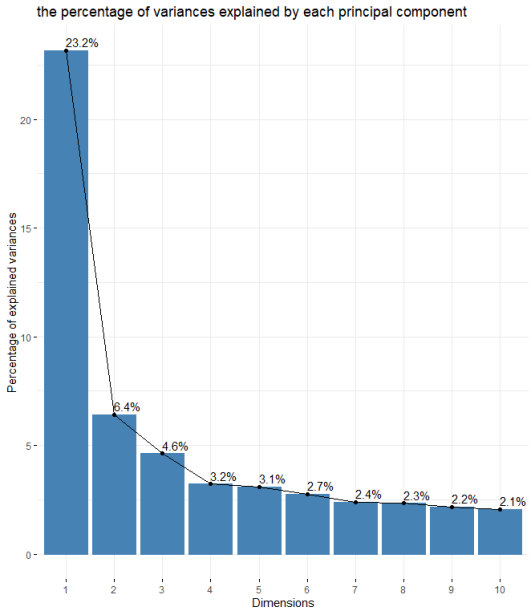
3.4 Position Switch Inference

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'.



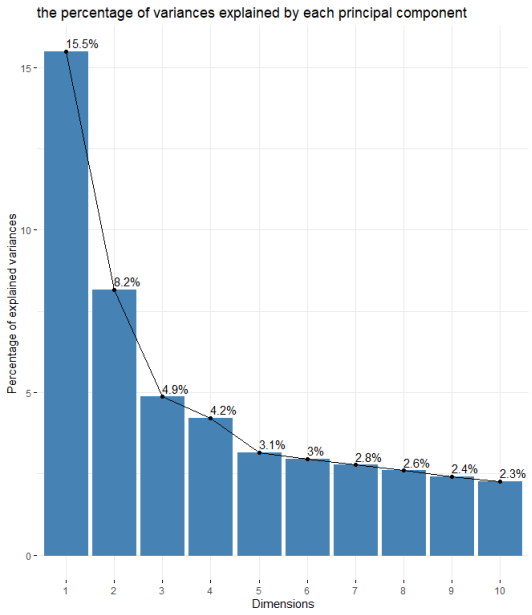
(a) PCA Graph for individuals



(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 Summary and Conclusions

4.1 Summary

From the current results of our projects, we discovered the following things:

1. The silent majority myth is true;
2. The distribution of the "swing voters" within each parties;
3. The variables that may help us identify these voters;
4. Tools for predicting the voting preference and the swing voter likelihood of voters;
5. We can predict the vote from the pre-review information.

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.

4.2 Future plan

1. Test for the stratification differences between the voters for DEM and the voters for GOP;
2. Adding more possible variables for the analysis;
3. Try to predict with only basic information of each people;
4. Detailed analysis about each label class of the political position switch;
5. Detailed voter image for each label class of the political position switch.