Optimus Interview Project

I used R to analyze the voter file by first importing it and then changed it to a data frame to make it easier to work with. Using the str() function allowed me to see the type or class of information for each variable or column. Then I could easily write the codes to convert all variables listed as characters to numerical values so that I could more easily run codes on the dataset. The summary() function is very helpful in the initial understanding of the dataset. In this case I saw that the variable of donating to liberal causes was completely blank, so it would be useless for further analysis of the dataset which is why I removed it. The summary function also showed me that the most NAs for any column was 17 out of 50,000 data points for Age. Knowing that I would still have over 90% of the 50,000 data points, I omitted all rows with NA in the dataset which only removed 22 out of 50,000 or 0.00044 of the data. Omitting NAs just allowed me to run the codes I needed later without any issues. This completed my cleanup of the dataset which was all done to make it easier to work with.

The first analysis I tried on the dataset was linear regression against vh14p. According to the linear regression test if alpha=0.10 the significant variables would be id, age, ethnicity, dwelling type, income, cd, dma, vh12g, vh12p, vh10p, vh08g, vh08p, vh06g, vh06p, vh04p, vh02p, vh00g, vh00p, donating to conservative causes, being a home owner or renter, g08 precinct turnout, p08 precinct turnout, and p12 precinct turnout which gave a total of 23 significant variables. However, I did not use this information to reduce the number of variables because the R^2=0.3662 which meant that there really was not much correlation between all the variables and vh14p. Next, I decided to use logistic regression on the dataset against vh14p. First, I had to remove the ID column because it should not be counted towards influencing the variable as simply a label. I could have used k-nearest neighbor or a classification tree, but I used logistic regression here because it is the one I am most comfortable with, and it seems to run the fastest on my R sessions.

For logistic regression I first partitioned the data with 60% in the training set and 40% in the testing set and then ran logistic regression on the training set. Once that was complete, I compared the predictions of the training set to the actual results of the testing set which is a good way to avoid overfitting. The comparison gave an error rate as well as a confusion matrix that clarified how well the logistic regression worked. Since the error rate was low, I used the logistic regression model to entirely predict the vh14p turnout probability after clearing its actual data. The reason I did not just use those results right away was because people may view primary elections and general elections differently, so I felt it was necessary to run it all again on a general election. After running logistic regression again on vh12g and getting the voter turnout probability, I then took the averages between every row of predicted vh14p and vh12g to then round that to predict whether the individuals would make it to the 2014 general elections. I could have run logistic regression models against the data for all the years of voter history for general and primary elections and then taken the averages from all those, but I decided to limit it to just vh14p and vh12g. The reason I only averaged the result of vh14p and vh12g is because those two were shown in the example provided as well as some elections were shown to not even be significant if we assumed alpha equaled 0.10.

The overall results predicted that there would be about a 12 to 13% voter turnout for the 2014 general elections. I got this prediction by keeping almost all the variables in the dataset except for the two aforementioned variables. I understand that it may be possible to get more accurate results by only keeping significant variables from a linear regression test, however I do believe that my results are accurate enough to predict voter turnout for the 2014 general election.