Classification tree prediction of self-reported voting using the General Social Survey

Ryan McCammon (Writer) and Michael Borsellino (Analyst)

Using polls to predict the outcome and margin in US presidential elections has a long and storied history. The utility of probability sampling was made clear to the public when the Gallup organization correctly predicted the outcome of the 1936 presidential election, in contrast to the widely accepted prediction of the Literary Digest based on a convenience sample of its readers (Crossley, 1937). As the election polling industry developed, the Gallup organization sought to refine its predictions by not only using probability samples, but also by adopting a "likely voter model" which seeks to identify those respondents most likely to vote in a given election (Gallup, 2017). The need for such a model in the prediction of presidential election outcomes is evident given that voter turnout averages less than 60% of the voting-age population (Peters & Wooley, 2017) and the fact that the electorate is not a random subset of this voting-age population.

There are seven questions used in the Gallup likely voter model, asking respondents about the thought that they have given to the election; knowledge of where to vote; whether they have voted in the precinct previously; how often they vote; whether or not they have a plan for voting in the upcoming election; self-reported likelihood of voting; and whether or not they voted in the last presidential election. While the likely voter model based on these items has been historically useful, its utility was called into question following Gallup's failure to correctly forecast the outcome of the 2012 presidential election (Gallup, 2013). The ability of pollsters to accurately identify the electorate was further called into question following the surprising outcome of the 2016 election (AAPOR, 2017).

Given the questionable performance of the Gallup likely voter model in the past two US presidential elections, our research question is whether a highly predictive model of voting behavior can be developed using a classification tree approach based on the recursive partitioning of survey responses to a variety of behavioral, attitudinal, and demographic items. We are hopeful that this approach will allow us to identify meaningful subgroups ("hypercubes") of the voting-age population that are relatively homogeneous in their voting behavior (voting/not voting).

Data set, analytic sample, and variable coding

The data for this project come from the 2006, 2010, and 2014 General Social Surveys (GSS) (Smith, et al, 2014). Using a repeated cross-sectional design, the GSS biennially surveys a probability sample of the US adult population on a variety of social, political, behavioral, and attitudinal, topics. In the first phase of the analysis, the 2014 GSS is used. In this survey, respondents self-report on their voting behavior in the 2012 and 2008 US elections. Self-reported voting in 2012 is the outcome of interest, and self-reported voting in 2008 is used to measure voting in the previous presidential election. Given the anticipated importance of prior voting behavior in predicting voting in the most recent presidential election, the 2014 GSS sample is restricted to respondents eligible to vote in the 2008 election, resulting in an analysis sample of 2,229 respondents aged 24 years and older in 2014.

Table 1. Unweighted descriptive statistics, 2014 GSS (n=2,229)

Respondent characteristics	mean(sd) or n(%)	Respondent characteristics	mean(sd) or n(%)
Voted in 2012 election		Social class	
no	644 (28.9)	lower class	192 (8.7)
yes	1585 (71.1)	working class	1004 (45.3)
Voted in 2008 election		middle class	958 (43.3)
no	572 (25.7)	upper class	61 (2.8)
yes	1657 (74.3)	Dwelling type	
Marital status		trailer	105 (5.1)
married	1069 (48.0)	detached 1-fam house	1369 (66.3)
widowed	199 (8.9)	duplex	87 (4.2)
divorced	393 (17.7)	3-4 fam house	24 (1.2)
separated	73 (3.3)	row house	90 (4.4)
never married	493 (22.1)	apartment	368 (17.8)
Religious preference		other	22 (1.1)
Protestant	1041 (47.0)	Labor force status	
Catholic	518 (23.4)	working full-time	1107 (49.7)
Jewish	36 (1.6)	working part-time	219 (9.8)
none	441 (19.9)	temporarily not working	36 (1.6)
other	181 (8.2)	unemployed, laid off	86 (3.9)
Geographic region		retired	446 (20.0)
New England	119 (5.3)	school	41 (1.8)
Middle Atlantic	276 (12.4)	keeping house	227 (10.2)
East north central	391 (17.5)	other	66 (3.0)
West north central	126 (5.7)	Highest educational degree	
South Atlantic	443 (19.9)	less than high school	265 (11.9)
East south central	131 (5.9)	high school	1095 (49.1)
West south central	232 (10.4)	junior college	168 (7.5)
Mountain	195 (8.8)	bachelor	443 (19.9)
Pacific	316 (14.2)	graduate	258 (11.6)
Race and ethnicity			
non-Hispanic white	1533 (68.9)	Age	51.2 (16.4)
non-Hispanic black	330 (14.8)	Inflation-adjusted family income	49901 (43340)
Hispanic	285 (12.8)	Frequency attend religious services	3.4 (2.8)
other	77 (3.5)	Political partisanship	1.7 (1.1)
Gender		Ideological intensity	1.0 (1.0)
male	990 (44.4)	Sample weight	0.97 (0.49)
female	1239 (55.6)		

Unweighted descriptive statistics for the voting behavior items, as well as for the other variables made available to the prediction algorithm, are found in Table 1. The text of the question used to measure the outcome variable of 2012 voting behavior is as follows, "In 2012, you remember that Obama ran for President on the Democratic ticket against Romney for the Republicans. Do you remember for sure whether or not you voted in that election?" A similar question was used to measure 2008 voting behavior, replacing "Romney" with "McCain" as the Republican candidate.

The variables in Table 1 for which frequencies and percentages are provided were treated as categorical (or "factor") variables, while unweighted means and standard deviations are presented for the continuous and ordinal measures. The mean value of 3.4 on the measure of frequency of attendance at religious services indicates a mean frequency of between "several times a year" and "once a month." Political partisanship is measured using a 4-point scale (0-3) where 0 refers to a political independent, and a value of 3 describes both "strong democrats" and "strong republicans." Ideological intensity is measured on a similar 4-point scale where 0 describes political moderates, and a value of 3 indicates those that consider themselves to be either "extremely liberal" or "extremely conservative" in their political views.

Use of statistical software

SAS software version 9.4 was used for all data management and coding, as well as for generating the descriptive statistics found in Table 1. The classification trees were built using the rpart package (Therneau, Atkinson, & Ripley, 2017) in R software version 4.3.2.

Modeling approach

We use classification trees to predict the binary outcome – voted/did not vote in the most recent presidential election. Classification trees are a data mining tool which seeks to partition cases into nodes that are as homogeneous as possible with respect to the outcome variable of interest. The construction of the classification tree begins with the partition (or "split") which results in the greatest reduction in misclassification within each resulting node. Each resulting node is then considered for further partitioning based on the next best split given the available predictors. Splits of a node are chosen based on the reduction of a deviance measure, D, defined as

$$D_v = -2\sum_{k=1}^{K} n_{vk} \log(P(v,k))$$

where v indicates the node, and k is the class (or outcome) assignment. In the case of the binary splitting of a node, the split selected will maximize the reduction in deviance, $D_{parent} - (D_{child1} + D_{child2})$. The final number of terminal nodes in a classification tree is determined by minimizing a complexity parameter that is a function of the size of the tree (# of terminal nodes) and the overall misclassification rate.

All variables found in Table 1 will be considered as potential splitting variables in the development of the classification tree. Missing data on any variable will be handled using the surrogate variable approach of the rpart package. This method replaces item missing data with the next best observed surrogate variable.

As a measure of the model's predictive power and a hedge against overfitting of the tree to idiosyncrasies of the training data, we will validate the classification tree through both cross-validation and by external validation – in this case, the application of the final tree developed using the 2014 GSS to external data from the 2010 and 2006 GSS datasets. To cross-validate our tree, we follow the default approach of rpart. This involves constructing ten trees based on 90% subsamples of the 2014 GSS and then applying each given tree to the remaining 10% of cases (the "out of bag" or OOB sample). The cross-validated error rate for the classification tree is then found by averaging the error rates for the ten OOB samples. The cross-validated error rate is an estimate of true error rate of the tree. This cross-validation procedure is also extremely useful in determining the appropriate size for the final tree. Using the results from the OOB sample, the cross-validated relative error of the tree can be observed as a function of the size of the tree (# of terminal nodes). Like a scree plot in factor analysis, a plot of the relative error by the number of terminal modes will be used to determine the optimal tree size.

In developing the tree, we first use a liberal criterion for the formation of new nodes, allowing a split to occur if the parent node has at least 10 cases and the reduction of deviance resulting from the split (the so-called "complexity parameter" (CP)) is as small as 0.001. By examining the relative error of this model as a function of tree size in the OOB samples, the model will be adjusted and re-estimated. In a final model, the GSS sample weight will be applied to the model. While the sample weight will be included as a potential splitting variable in the development of the tree, using it to weight each case directly will allow for better assessment of the sensitivity of the results to the complex sample design.

Finally, to gauge the external validity of the classification tree, it will be applied to the 2010 GSS to predict 2008 voting behavior, and to the 2006 GSS to predict the self-report of voting in the 2004 presidential election.

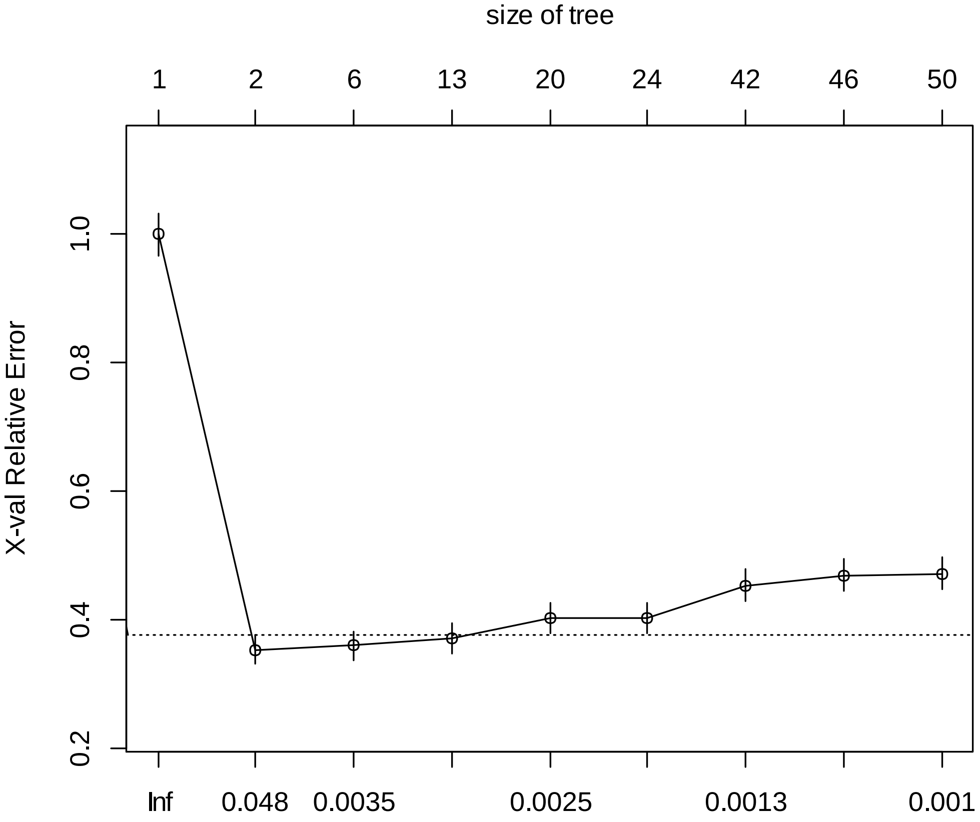
Results

The first classification tree, developed using liberal splitting rules (CP=0.001) found many splits in the data that resulted in a reduction of overall model deviance. Variables important in this model include previous voting behavior, educational degree, geographic region, age, race/ethnicity, dwelling type, and many others. The relative and absolute classification error rates for this tree are summarized in Table 2 and Figure 1.

Table 2. 2012 voting behavior: classification tree error rates by the number of splits (2014 GSS, n=2,229)

	relative error			error rate	
# of splits	training sample	OOB sample mean	OOB sample sd	apparent	true
0	1.000	1.000	0.033	0.289	0.289
1	0.354	0.354	0.022	0.102	0.102
5	0.340	0.360	0.022	0.098	0.104
12	0.315	0.371	0.023	0.091	0.107
19	0.293	0.402	0.023	0.085	0.116
23	0.286	0.404	0.024	0.083	0.117
41	0.258	0.453	0.025	0.074	0.131
45	0.253	0.469	0.025	0.073	0.135
49	0.248	0.472	0.025	0.072	0.136

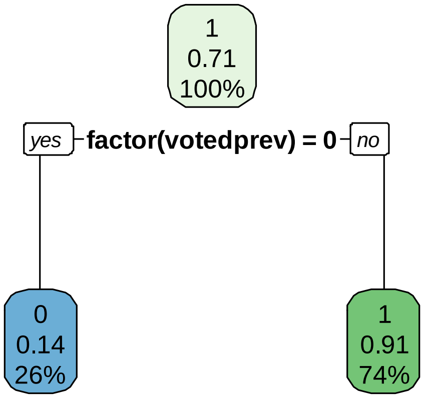
Figure 1. Cross-validated relative error rate as a function of tree size



With no splits (1 node), the apparent and estimated true error rates are the same, 0.289. In this single node tree, all cases are classified as voters, and the error rate is simply the proportion of self-reported non-voters in the 2012 election. In a tree with 49 splits, the apparent error rate is reduced to 0.072, with an estimated true error rate based on the cross-validated OOB samples of 0.136. Importantly, however, as is shown in Figure 1, the cross-validated relative error rate is minimized with just one split (size of tree=2).

On the basis of these initial results, we generate a new tree, restricting the size of the tree to just two nodes. Under this model, there is just one split, defined by previous voting behavior (voting in the 2008 election). A visual representation of this tree is shown in Figure 2.

Figure 2. 2012 voting behavior: Final classification tree (2014 GSS)



In this tree, the blue node contains those cases classified as non-voters in 2012. This node contains all respondents who self-reported not voting in the 2008 election. 26% of the sample falls into this node, and 14% of these cases self-report having voted in the 2012 election. The green node contains those cases classified as voters in 2012. This node contains all respondents who self-reported voting in the 2008 election. 74% of the sample falls into this node, and 91% of these cases self-report having voted in the 2012 election. Based on cross-validation, the estimated true error rate for this classification tree is 10.2%.

Weighting

The GSS sample weight WTSSALL was included as a potential splitting variable in the models presented above. This weight primarily adjusts for a two-stage subsampling design used to correct for nonresponse. Since the weight was not used for any splits in the final tree, this suggests that the probability of voting is not related to response propensity in the 2014 GSS. As an additional test of the effect of weighting, the tree was re-estimated using rpart and applying WTSSALL as a case weight. Weighting the data had no meaningful effect on the results. A two-node tree was still selected, and the cross-validated relative error increased trivially from 0.354 to 0.369.

External validation

As a test of the external validity of the final classification tree, the prediction model was applied to data from the 2010 and 2006 GSS. The confusion matrices for the training data (2014 GSS) as well as for these two external samples are found in Table 3.

Table 3. Confusion matrices, GSS 2006, 2010, and 2014

	2014 GSS				2010 GSS				2006 GSS		
	Predicted				Predicted				Predicted		
Observed	non-voter	voter	Total	Observed	non-voter	voter	Total	Observed	non-voter	voter	Total
non-voter	546	98	644	non-voter	326	121	447	non-voter	485	143	628
voter	62	1523	1585	voter	198	1088	1286	voter	246	1640	1886
Total	608	1621	2229	Total	524	1209	1733	Total	731	1783	2514

The apparent error rate in the training data is 0.072 ((62 + 98) / 2229), while estimates of the true error rate based on external data are 0.184 (2010 GSS) and 0.155 (2006 GSS).

Discussion

We hoped that the use of classification tree methods would allow for the identification of complex interactions that would improve on the prediction of voting behavior, our analysis failed to identify any meaningful predictors of voting beyond previous voting behavior. While this finding may reflect the lack of suitable predictors in the GSS, it also suggests that identifying likely voters is a challenging task, as is known all too well by the political polling firms. Not surprisingly, past behavior is predictive of future behavior, and voting may well be understood as a habitual behavior, a view which is consistent with prior research (Fowler, 2006).

An important caveat to this analysis is that the data are self-reported and were collected after the election of interest. It is well-known that voting is over-reported by survey respondents (Bernstein, Chadha, & Montjoy, 2001), and the reduction in response variance caused by this phenomenon may make it more difficult to identify good predictors. Additionally, the GSS questions about voting in the immediately past and previous elections were asked in close succession in the survey. This may result in a consistency bias which results in an overestimation of the importance of previous voting in the prediction of more recent voting behavior. Additional work using classification trees to predict voting should be done in samples with a record check of voting activity, rather than relying on self-report. Ideally, such data would include predictors more specifically aimed at distinguishing voters from non-voters.

References

AAPOR. An evaluation of 2016 election polls in the United States. 2017.

Bernstein R, Chadha A, Montjoy R. Overreporting voting: Why it happens and why it matters.

Public Opinion Quarterly. 2001;65(1):22-44.

Crossley AM. Straw polls in 1936. Public Opinion Quarterly. 1937;1(1):24-35.

Fowler JH. Habitual voting and behavioral turnout. Journal of Politics. 2006;68(2):335-44.

Gallup. Gallup 2012 Presidential Election Polling Review. 2013.

Gallup. Understanding Gallup's Likely Voter Procedures for Presidential Elections: Gallup Corporation; 2017 [Available from: http://news.gallup.com/poll/111268/how-gallups-likely-voter-models-work.aspx.

Peters G, Woolley J. Voter Turnout in Presidential Elections: 1828 – 2012. 2017 [Available from: http://www.presidency.ucsb.edu/data/turnout.php.

Smith, Tom W, Peter Marsden, Michael Hout, and Jibum Kim. General Social Surveys, 1972-2014 [machine-readable data file] /Principal Investigator, Tom W. Smith; Co-Principal Investigator, Peter V. Marsden; Co-Principal Investigator, Michael Hout; Sponsored by National Science Foundation. -NORC ed.- Chicago: NORC at the University of Chicago [producer and distributor].

Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1–11. https://CRANR-projectorg/package=rpart 2017.

Appendix

SAS code: data management and recoding

```
libname gss 'F:\DATA\GSS';
libname cart 'F:\mborsel\ANALYSIS\PSM\SurvMeth746\CART';

data work.gss;
    set gss.GSS7216_R2;

    /* estimate date of birth */

    array bmonth{12} _temporary_ (4 5 6 7 8 9 10 11 12 1 2 3);
    array bday{12} _temporary_ (19 20 20 22 22 22 22 21 21 19 18 20);
    if cohort ge 1883 and cohort le 1999 then do;
        do lcv=1 to 12;
            if lcv eq zodiac then dob = mdy(bmonth{lcv},bday{lcv},cohort);
        end;
        if dob eq . then dob = mdy(12,31,cohort);
    end;
    eday2012 = mdy(11,6,2012);
    eday2008 = mdy(11,4,2008);
    eday2004 = mdy(11,2,2004);
    eday2000 = mdy(11,7,2000);

    if dob gt .z then do;
        if yrdif(dob,eday2012) ge 18 then ageElig2012 = 1; else ageElig2012=0;
        if yrdif(dob,eday2008) ge 18 then ageElig2008 = 1; else ageElig2008=0;
        if yrdif(dob,eday2004) ge 18 then ageElig2004 = 1; else ageElig2004=0;
        if yrdif(dob,eday2000) ge 18 then ageElig2000 = 1; else ageElig2000=0;
    end;

    /* collapse "other" religions */
    if relig in (6 7 8 9 10 11 12 13) then relig = 5;

    /*race ethnicity */
    if hispanic ge 2 and hispanic le 50 then raceeth = 3;
    else if hispanic eq 1 and race eq 1 then raceeth = 1;
    else if hispanic eq 1 and race eq 2 then raceeth = 2;
    else if hispanic eq 1 and race eq 3 then raceeth = 4;

    /* dwelling: duplex and apartment */
    if dwelling in (3 4) then dwelling = 3;
    else if dwelling in (7 8 9) then dwelling = 7;



/* R's that self-report as ineligible to vote are set missing */

if vote12 eq 1 then voted12 = 1;
else if vote12 eq 2 then voted12 = 0;

if vote08 eq 1 then voted08 = 1;
```

```
else if vote08 eq 2 then voted08 = 0;

if vote04 eq 1 then voted04 = 1;
else if vote04 eq 2 then voted04 = 0;

if vote00 eq 1 then voted00 = 1;
else if vote00 eq 2 then voted00 = 0;

if partyid eq 3 then partisan = 0;
else if partyid in (2 4) then partisan = 1;
else if partyid in (1 5) then partisan = 2;
else if partyid in (0 6) then partisan = 3;

if polviews eq 4 then intensity = 0;
else if polviews in (3 5) then intensity = 1;
else if polviews in (2 6) then intensity = 2;
else if polviews in (1 7) then intensity = 3;

data cart.gss2014 (keep=id year voteprev votedprev vote voted partyid polviews partisan intensity
                marital relig region raceeth sex class age dwelling wrkstat degree income06 attend coninc
ageElig ageEligPrev WTSSALL);
                set work.gss;
                if year eq 2014 and voted12 in (0 1) and ageElig2012 eq 1 and voted08 in (0 1) and
ageElig2008 eq 1;

                voteprev = vote08;
                votedprev = voted08;
                ageEligPrev = ageElig2008;
                vote = vote12;
                voted = voted12;
                ageElig = ageElig2012;


data cart.gss2010 (keep=id year voteprev votedprev vote voted partyid polviews partisan intensity
                marital relig region raceeth sex class age dwelling wrkstat degree income06 attend coninc
ageElig ageEligPrev WTSSALL);
                set work.gss;
                if year eq 2010 and voted08 in (0 1) and ageElig2008 eq 1 and voted04 in (0 1) and
ageElig2004 eq 1;

        voteprev = vote04;
        votedprev = voted04;
        ageEligPrev = ageElig2004;
        vote = vote08;
        voted = voted08;
        ageElig = ageElig2008;

data cart.gss2006 (keep=id year voteprev votedprev vote voted partyid polviews partisan intensity
        marital relig region raceeth sex class age dwelling wrkstat degree income06 attend coninc ageElig
        ageEligPrev WTSSALL);
        set work.gss;
        if year eq 2006 and voted04 in (0 1) and ageElig2004 eq 1 and voted00 in (0 1) and ageElig2000 eq 1;
        voteprev = vote00;
        votedprev = voted00;
        ageEligPrev = ageElig2000;
        vote = vote04;
        voted = voted04;
        ageElig = ageElig2004;
run;

proc freq data=cart.gss2014;
        tables voted votedprev marital relig region raceeth sex class dwelling wrkstat degree partisan intensity;
run;

proc univariate data=cart.gss2014;
        var coninc attend partisan intensity age wtssall;
run;
```

R code: Classification tree

```
sink("F:/mborsel/ANALYSIS/PSM/SurvMeth746/CART/gssVoteOutput.txt", append=FALSE,
split=TRUE)

library(rpart)
library(rpart.plot)
library(descr)

# build tree using 2014 GSS
load("F:/mborsel/ANALYSIS/PSM/SurvMeth746/CART/gss2014.Rdata")

# outcome variable is voted in 2012 election

# relax complexity parameter constraint to look for all useful splits
g.control <- rpart.control(minsplit=10, cp=0.001)
tree1 <- rpart(voted ~ factor(votedprev) + factor(marital) + factor(relig) + factor(region)
+ factor(raceeth)
          + factor(sex) + factor(class) + factor(dwelling) + factor(wrkstat) + factor(degree) + coninc
          + attend + partisan + intensity + age + wtssall, data = gss2014, method =
"class",control=g.control)
printcp(tree1)
plotcp(tree1)

g.control <- rpart.control(minsplit=10, cp=0.048)
tree2 <- rpart(voted ~  factor(votedprev) + factor(marital) + factor(relig) + factor(region) + factor(raceeth)
          + factor(sex) + factor(class) + factor(dwelling) + factor(wrkstat) + factor(degree) + coninc
          + attend + partisan + intensity + age + wtssall, data = gss2014, method = "class")
printcp(tree2)
plotcp(tree2)
summary(tree2)
rpart.plot(tree2)

#treat sample weight as WEIGHT rather than using as a predictor
tree3 <- rpart(voted ~ factor(votedprev) + factor(marital) + factor(relig) + factor(region) + factor(raceeth)
        + factor(sex) + factor(class) + factor(dwelling) + factor(wrkstat) + factor(degree) + coninc
        + attend + partisan + intensity + age, data = gss2014, weights = wtssall, method = "class")
printcp(tree3)
plotcp(tree3)
summary(tree3)
rpart.plot(tree3)

#training data confusion matrix
tree.pred2014<-predict(tree1,gss2014,type="class")
CrossTable(gss2014$voted,tree.pred2014,prop.chisq = FALSE)

#external validation of tree using 2010 and 2006 GSS
load("F:/mborsel/ANALYSIS/PSM/SurvMeth746/CART/gss2010.Rdata")
tree.pred2010<-predict(tree1,gss2010,type="class")
CrossTable(gss2010$voted,tree.pred2010,prop.chisq = FALSE)

load("F:/mborsel/ANALYSIS/PSM/SurvMeth746/CART/gss2006.Rdata")
tree.pred2006<-predict(tree1,gss2006,type="class")
CrossTable(gss2006$voted,tree.pred2006,prop.chisq = FALSE)
```