# Predicting the 2000 USA Presidential Election Results at the County Level Using Public Data

Mose Wintner

## 1 Problem

How do citizens of a democratic, or for that matter, of any society come to form political opinions? It seems sensible to hypothesize that an individual's position on a given political issue is likely to agree with the opinion at large in his or her community, if there is one. By the same token, the history of a community determines its demographics and informs its present-day culture. Thus the demographics of a community are likely to, at least sometimes, be an unreliable proxy for its history and, by extension, its culture, and by further extension, its politics.

I set out to test this speculation in the context of the 2000 U.S. presidential election, with county-level data. I was interested in testing, for a few different models, which variables were most important in determining election results. I trained regression models on `rpct`, the percent of a given county that voted for Bush in 2000; and classification models on `rpct` $> 50$, a binary variable indicating who won the county vote. I decided to use numbers for Bush because I am most interested in measures of political opinion at large, and the most popular third-party candidate by far in 2000 was Ralph Nader, who was closer politically to Gore. Thus regressions on the proportion of the county voting Bush would be a more reliable regression of left/right ideologies at large than would regressions on the proportion voting Gore. Ideal would have been finer demographics, since a county is probably too large to be considered a community, that is, a unit with a cohesive enough culture to indicate its politics. Perhaps such data is more readily available today.

One could conceivably expand upon these results to use panel data, that is, perform this task for the same variables, for other presidential elections, and explore trends and variations. From such panel data and analysis, one could attempt to make predictive models for upcoming elections, especially if demographic data is available on a fine enough geographic scale for each observation to approximate well a "community".

This project was exploratory in nature and did not put forth any hypotheses, except the one hypothesis per model that demographics have nonzero predictive power in predicting `rpct`, where "predictive power" is model-specific. Of course, we may regard $R^2$ as a random variable depending on the data and model, which relates to the aforementioned "predictive power". I used this freedom to test many different models. To cross-validate and evaluate models, I used the R package `caret`, a versatile data analysis R package.

## 2 Data

Almost all the data came from the U.S. Census Bureau's Counties Database [1], which is no longer actively maintained. After cleaning the data and removing data for Alaska, which evidently suffers from inconsistent districting, I used data on $n = 3074$ counties. After running a correlation analysis and engineering some features, I ended up with $p = 77$ total variables consisting of and engineered from demographic and housing data. An exhaustive list of the predictors and corresponding descriptions can be found at the end. All data is from 2000 unless otherwise noted.

## 3 Classification

For the classification problem, I trained several models: SVM, LDA, QDA, KNN, and Adaboost. 73.3% of counties voted for Bush in 2000. Model parameters were found via grid search, by tenfold cross-validation. Then a separate repeated tenfold cross-validation was carried out on these models. Here is a table of the average results on the out-of-bag portion of the data during cross-validation for these optimal models.

| Method | Parameters | Mean Accuracy | Mean Kappa |
|---|---|---|---|
| SVM | C=1 | 0.8389568 | 0.5554758 |
| LDA | N/A | 0.8345383 | 0.5330159 |
| QDA | N/A | 0.8125396 | 0.4705189 |
| KNN | k=30 | 0.7475475 | 0.1237399 |
| Adaboost | iter=250, maxdepth=6 | 0.8584849 | 0.6061885 |

Cohen's kappa is $(p_0 - p_e)/(1 - p_e)$, where $p_0$ is the average relative observed agreement of the cross-validated models on out-of-bag data, and $p_e$ is the hypothetical probability of chance agreement of the cross-validated models on the out-of-bag data based on the prior distribution.

The SVM parameter indicates the cost of misclassification of a single point. LDA and QDA performed better than expected, since among the predictors there are many heavy-tailed distributions and these algorithms assume Gaussian priors. For this reason, they were perhaps inappropriate. I considered using a Box-Cox or Yeo-Johnson transformation on these heavy-tailed distributions, but decided against it. Adaboost is a relatively sophisticated tree-based boosting algorithm based on minimizing a weighted misclassification error. iter selects the number of "sequential" trees to train, and maxdepth is the depth of each of these trees. We see three of the five methods hovering around 84% accuracy, which is a relatively strong result. I considered examining the misclassified results, but I wasn't really so interested in the classification model. The 50% threshold is arbitrary, at least for my purposes.

## 4 Regression on `rpct`

I was most interested in prediction and variable importance measures for the regression problem. The average vote for Bush in counties in the U.S. was 56.9%. I used tenfold cross-validation to determine optimal model parameters by grid search. Then, repeated tenfold cross-validations were used on the optimal models. The results are summarized below.

| Method | Parameters | Rsquared | RsquaredSD |
|---|---|---|---|
| Lasso | $\lambda = 0.110$ | 0.6415329 | 0.05642351 |
| Random Forest | mtry=30 | 0.7334688 | 0.03450272 |
| SVM | C=0.5 | 0.6481942 | 0.05012503 |
| GBM | n.trees=450, depth=11, n.minobsinnode=20 | 0.7721142 | 0.03053278 |
| Bagged CART | N/A | 0.5646940 | 0.02516780 |
| XGBLinear | nrounds=250 | 0.7405678 | 0.02932354 |
| XGBTree | nrounds=200, maxdepth=4 | 0.7478437 | 0.02707445 |

The lasso parameter is a shrinkage parameter. The random forest parameter indicates the number of features looked at when determining each split. The SVM parameter is a weight for the least squares error function. GBM is stochastic gradient boosting, developed by Friedman. One trains a sequence of trees, as with the ordinary gradient boosting tree algorithm. However, we require that there be at most $L$ terminal nodes per tree, and each tree is trained only on a randomly selected subsample of training data. This mitigates overfitting and generally performs better than random forest. The parameter n.trees controls the number of trees to sequence, depth fixes the maximum number of leaves of each tree, and the algorithm will not split nodes containing fewer than n.minobsinnode observations. Bagged CART is an ordinary tree bagging algorithm. XGBoost was developed by Chen & Guestrin, and can be described as a regularized stochastic gradient boosting algorithm. It is usually used for big data, because the algorithm is distributed, generally works very well on large datasets, and sparse data-aware.

We see from our models that we could expect demographics to generally explain just over 70% of the variation in 2000, which is more than I had expected. The most successful model, by one standard error, was stochastic gradient boosting, which explained 77.2% of a county vote's deviation from the mean vote (for Bush).

But surely some predictors were weak in predicting how a county voted. Which ones? This is a variable selection problem. However, since there are 77 variables, even a forward selection algorithm to find a best 5-variable model would take a very long time, since $\sum_{k=1}^{5} \binom{77}{k} \approx 2 \times 10^7$. Instead, we used measures of variable importance included in `caret`. For linear models, variable importance for a given predictor is taken as the absolute value of its $t$-statistic.

For the random forest's variable importance, the MSE is computed on the out-of-bag data for each tree, and then the same computed after permuting a variable. The differences are averaged and normalized by the standard error. If the standard error is equal to 0 for a variable, the division is not done.

For recursive partitioning methods, "[t]he reduction in the loss function (e.g. mean squared error) attributed to each variable at each split is tabulated and the sum is returned. Also, since there may be candidate variables that are important but are not used in a split, the top competing variables are also tabulated at each split."

For boosted or bagged tree algorithms, the same method is used as for a single tree, with MSE computed over all bootstrapped trees.

All variable importance measures were then scaled to range 0-100; this does not change the relative magnitudes of the importance values. I decided to first explore the lasso as a linear variable selection method. A positive coefficient indicates the variable is associated with a higher proportion of the county's vote going to Bush, negative, for Gore.

Lasso, $\lambda = 0.110$

| variable | scaled $\|t\|$ | coefficient |
|---|---|---|
| `married` | 100.00000 | 3.982 |
| `white1nh` | 90.73494 | 3.597 |
| `latitude` | 79.23087 | -3.136 |
| `vehperhouse` | 58.08542 | 2.311 |
| `voterparticip` | 42.87387 | -1.705 |
| `longitude` | 42.33608 | 1.668 |
| `nohealthinsurance` | 38.40696 | 1.515 |
| `lnpoppersqm` | 35.62778 | -1.465 |
| `mediangrossrent` | 35.12867 | -1.369 |
| `hisppop` | 30.43560 | -1.151 |
| `samehouse` | 29.85393 | -1.190 |
| `age45_54` | 27.37211 | -1.091 |
| `unemployedclf` | 26.20114 | -1.045 |
| `gwagegap` | 25.28031 | 0.995 |
| `other1` | 24.25983 | 0.931 |
| `manprofoccs` | 23.78744 | 0.917 |
| `multrace` | 22.90927 | -0.908 |
| `european` | 22.22989 | -0.888 |
| `grad` | 22.13447 | -0.823 |
| `medianhvalue` | 21.67482 | -0.861 |

$\lambda$ was selected by the one SE method, meaning the greatest value of $\lambda$ was chosen among those tested so that the associated $R^2$ was within one standard error of that of the full linear model. For this value of $\lambda$, the following 26 variables were eliminated: `emplgov`, `hospinsured`, `publicindus`, `publicwaterpercap`, `irrigationwater`, `onfarms`, `black1`, `nevermarried`, `widowed`, `foreignborn`, `naturalized`, `samecounty`, `veteran`, `salesoffoccs`, `consoccs`, `poor`, `mobilehomes`, `avgageunit`, `nocashrent`, `lnpop`, `sepdiv`, `housingincomediscrep`, `over55`, `age18to35`, `asian`, `bankspер1000pop`.

Then I used `caret` to calculate variable importances for the other methods. Below are listed the top 10 for each method

| Random Forest | |
|---|---|
| `married` | 100.00000 |
| `latitude` | 89.89273 |
| `white1nh` | 74.55183 |
| `lnpoppersqm` | 67.49755 |
| `longitude` | 65.79855 |
| `voterparticip` | 64.61488 |
| `european` | 46.66031 |
| `irrigationwater` | 45.50158 |
| `vehperhouse` | 44.30524 |
| `black1` | 43.87877 |

| SVM | |
|---|---|
| `married` | 100.00000 |
| `vehperhouse` | 79.80770 |
| `nevermarried` | 69.66694 |
| `lnpoppersqm` | 63.34632 |
| `lnpop` | 50.25822 |
| `white1nh` | 45.73674 |
| `publicindus` | 40.53753 |
| `onfarms` | 37.97240 |
| `unemployedclf` | 35.03867 |
| `black1` | 31.25504 |

| Bagged CART | |
|---|---|
| `married` | 100.00000 |
| `vehperhouse` | 89.18799 |
| `white1nh` | 75.76427 |
| `lnpoppersqm` | 67.69912 |
| `mobilehomes` | 50.83137 |
| `irrigationwater` | 46.27078 |
| `farmfishoccs` | 39.91635 |
| `nevermarried` | 38.41166 |
| `european` | 33.18302 |
| `voterparticip` | 27.54177 |

| Stochastic Gradient Boosting | |
|---|---|
| `married` | 100.000000 |
| `white1nh` | 30.195746 |
| `lnpoppersqm` | 29.143112 |
| `latitude` | 21.147230 |
| `longitude` | 20.295298 |
| `vehperhouse` | 16.882951 |
| `european` | 12.616580 |
| `voterparticip` | 8.606195 |
| `irrigationwater` | 8.403941 |
| `mobilehomes` | 7.442040 |

| XGBoost Linear | |
|---|---|
| `married` | 100.000000 |
| `lnpoppersqm` | 36.208238 |
| `white1nh` | 25.779966 |
| `vehperhouse` | 20.182694 |
| `european` | 15.525901 |
| `longitude` | 15.002006 |
| `latitude` | 14.585078 |
| `domes..ercap` | 9.377486 |
| `black1` | 5.702451 |
| `voterparticip` | 5.146734 |

| XGBoost Trees | |
|---|---|
| `married` | 100.000000 |
| `vehperhouse` | 28.415682 |
| `lnpoppersqm` | 19.172165 |
| `white1nh` | 16.947449 |
| `longitude` | 14.699382 |
| `european` | 11.990038 |
| `latitude` | 9.642352 |
| `lnland` | 8.979136 |
| `nevermarried` | 7.947897 |
| `black1` | 7.568393 |

We see that the variable indicating what percent of a county was married in 2000 was invariably the best predictor in determining a county's vote in the 2000 presidential election, especially so for stochastic gradient boosting, which performed best. To obtain a final score for each variable, I scaled the $R^2$ values to have range 0-1. I then weighted importance scores by their scaled $R^2$ values and summed the scores for each variable. Note that this effectively drops the least successful method. The twenty most important variables by this measure are below.

Sum of scores weighted by scaled $R^2$

| | |
|---|---|
| `married` | 431.79529 |
| `white1nh` | 194.74779 |
| `lnpoppersqm` | 193.42342 |
| `vehperhouse` | 171.51529 |
| `latitude` | 148.61230 |
| `longitude` | 129.50990 |
| `european` | 105.01412 |
| `nevermarried` | 98.29973 |
| `voterparticip` | 92.84896 |
| `black1` | 77.29452 |
| `domesticwaterpercap` | 70.24794 |
| `unemployedclf` | 69.44665 |
| `farmfishoccs` | 64.97608 |
| `irrigationwater` | 59.62696 |
| `nohealthinsurance` | 58.69887 |
| `lnpop` | 55.16021 |
| `medianhvalue` | 54.22515 |
| `onfarms` | 54.10720 |
| `pci` | 52.30743 |
| `publicindus` | 50.06372 |

This is the first time we have made a statistic from quantities which should not perhaps be compared to one another, namely, the variable importance scores. Still, it does give some illustration of the most important predictors. I would not have expected married nor vehperhouse to score so highly.

# 5 Regression on `rvotes`

I also figured it would be interesting to investigate regressions on the actual vote count rather than the percentage. Obviously, `lnpop` is the most significant predictor, but what about beyond that? Most appropriate for a regression on a count variable is usually a Poisson regression, the full model of which attained a mean $R^2 = 0.5367002$ on out-of-bag data for repeated tenfold cross-validation.

Somewhat less appropriate are regressions on `ln(rvotes)`, the natural logarithm of the number of votes. In such regressions, errors are measured in *orders of magnitude*. Thus the resulting $R^2$ measures, roughly, proportion of the *variation in orders of magnitude* explained, which is not so easily interpretable. Linear methods are generally regarded as inappropriate for count data. Thus we use the tree-based regression methods we used previously on the the logarithm of the Bush vote count.

Here are the cross-validation results.

| `ln(rvotes)` | Random Forest | GBM | Bagged CART | XGBTree |
|---|---|---|---|---|
| $R^2$ mean | 0.87522861 | 0.88530812 | 0.82228664 | 0.86760423 |
| $R^2$ SD | 0.07205758 | 0.04974909 | 0.02805907 | 0.06211919 |

Random Forest

| | |
|---|---|
| `lnpop` | 100.00000 |
| `lnpoppersqm` | 36.95073 |
| `white1nh` | 34.82902 |
| `voterparticip` | 27.49469 |
| `urban` | 25.90392 |
| `lnland` | 24.53248 |
| `married` | 24.11475 |
| `latitude` | 23.03701 |
| `latinamerican` | 22.22681 |
| `flow10yr` | 21.54858 |

GBM

| | |
|---|---|
| `lnpop` | 100.0000000 |
| `longitude` | 4.4341801 |
| `latitude` | 2.9070843 |
| `white1nh` | 1.9125250 |
| `publicindus` | 1.4249750 |
| `voterparticip` | 1.3313717 |
| `industrialwater` | 1.1279220 |
| `domesticwaterpercap` | 1.1151771 |
| `somecollege` | 0.9787950 |
| `latinamerican` | 0.9371652 |

Bagged CART

| | |
|---|---|
| `lnpop` | 100.000000 |
| `lnpoppersqm` | 60.270774 |
| `urban` | 36.532274 |
| `farmfishoccs` | 27.145071 |
| `nocashrent` | 20.756185 |
| `onfarms` | 14.419729 |
| `medianhvalue` | 9.920545 |
| `salesoffoccs` | 6.890640 |
| `latitude` | 6.757325 |
| `mediangrossrent` | 5.245729 |

XGBTree

| | |
|---|---|
| `lnpop` | 100.000000 |
| `lnpoppersqm` | 21.691523 |
| `nocashrent` | 17.647713 |
| `medianhvalue` | 4.558348 |
| `lnland` | 3.175129 |
| `latitude` | 3.030779 |
| `asian` | 2.115244 |
| `domesticwaterpercap` | 1.632391 |
| `transoccs` | 1.241441 |
| `emplstatelocalgov` | 1.007092 |

Score

| | |
|---|---|
| `lnpop` | 252.470084 |
| `lnpoppersqm` | 76.981793 |
| `urban` | 49.009299 |
| `farmfishoccs` | 34.058159 |
| `nocashrent` | 29.658176 |
| `white1nh` | 26.636208 |
| `latitude` | 23.821280 |
| `onfarms` | 23.754884 |
| `medianhvalue` | 22.870356 |
| `voterparticip` | 19.486242 |

These variables seem to conform more with what my intuition was before conducting this project.

On the other hand, one advantage of declining to transform the response, that is, predicting the response `rvotes`, is a more interpretable $R^2$. Another is that voting patterns associated with some predictor values lying in the tail of one or more heavy-tailed distributions are more likely to be identified. The correlation of `rvotes` and the population of each county is 0.74, which is lower than I would have expected. For this regression, I used the actual population rather than its logarithm. Here are the cross-validation results.

| | Poisson | Random Forest | GBM | Bagged CART | XGBTree |
|---|---|---|---|---|---|
| Rsquared | 0.5367781 | 0.6415233 | 0.6528643 | 0.6663007 | 0.6092250 |
| RsquaredSD | 0.1765128 | 0.2369861 | 0.2574123 | 0.2387208 | 0.2356744 |

Score

| | |
|---|---|
| `lnpop` | 277.251032 |
| `lnpoppersqm` | 114.652632 |
| `urban` | 82.456947 |
| `latinamerican` | 82.208369 |
| `samehouse` | 81.546792 |
| `voterparticip` | 81.218069 |
| `transoccs` | 73.314159 |
| `samecounty` | 67.999533 |
| `nocashrent` | 66.411812 |
| `lnland` | 65.284508 |

It's not surprising that this regression was less accurate, since the response is exponentially distributed. I am, however, surprised that `lnpop` was not far more important than the rest.

# 6 Appendix

| | names | description |
|---|---|---|
| 1 | emplgov | Percent of county employed by government |
| 2 | emplstatelocalgov | Percent of county employed by state and local government |
| 3 | hospinsured | Percent of county enrolled in hospital insurance or Medicare |
| 4 | nohealthinsurance | Percent of county without health insurance coverage |
| 5 | vehperhouse | Average number of vehicles per household |
| 6 | publicindus | Percent of county employed in public administration |
| 7 | flow10yr | Percent change in population 1990-2000 |
| 8 | birthsper1000 | Birth rate per 1000 population |
| 9 | deathsper1000 | Death rate per 1000 population |
| 10 | publicwaterpercap | Public water usage per capita |
| 11 | domesticwaterpercap | Domestic water usage per capita |
| 12 | irrigationwater | Total irrigation water usage |
| 13 | industrialwater | Total industrial water usage |
| 14 | onfarms | Percent of county living on farms |
| 15 | age35_44 | Percent of county aged 35-44 |
| 16 | black1 | Percent of county self-reported race as black only |
| 17 | indian1 | Percent of county self-reported race as Native American only |
| 18 | asian1 | Percent of county self-reported race as Asian only |
| 19 | other1 | Percent of county self-reported race category as 'other' only |
| 20 | multrace | Percent of county self-reported as of two or more races |
| 21 | hisppop | Percent of county self-reported as Hispanic |
| 22 | white1nh | Percent of county self-reported as white only and non-Hispanic |
| 23 | nevermarried | Percent of county that has never married |
| 24 | married | Percent of county currently married |
| 25 | widowed | Percent of county widowed |
| 26 | englishonly | Percent of county that speaks English only |
| 27 | noenglish | Percent of county that speak no English |
| 28 | foreignborn | Percent of county that is foreign-born |
| 29 | naturalized | Percent of county that are naturalized citizens |
| 30 | enteredlast10yrs | Percent of county that entered the country in the last 10 years |
| 31 | samehouse | Percent of county that lived in the same house in 1995 |
| 32 | samecounty | Percent of county that lived in the same county in 1995 |
| 33 | inpvtschools | Percent of county currently attending private school |
| 34 | incollege | Percent of county currently in college |
| 35 | lessthan9th | Percent of county whose highest educational attainment was less than 9th grade |
| 36 | somehighschool | Percent of county whose highest educational attainment was some high school |
| 37 | highschool | Percent of county whose highest educational attainment was a high school diploma |
| 38 | somecollege | Percent of county whose highest educational attainment was some college |
| 39 | bachelors | Percent of county whose highest educational attainment was a bachelor's degree |
| 40 | veteran | Percent of county who are military veterans |
| 41 | civlabforce | Percent of county in the civilian labor force |
| 42 | unemployedclf | Percent of county civilian labor force who are unemployed |
| 43 | manprofoccs | Percent of occupations in county that are management or professional |
| 44 | serviceoccs | Percent of occupations in county in service |
| 45 | salesoffoccs | Percent of occupations in county in sales or office |
| 46 | farmfishoccs | Percent of occupations in county in farming, fishing, and other agriculture |
| 47 | consoccs | Percent of occupations in county in construction |
| 48 | transoccs | Percent of occupations in county in transportation |
| 49 | pci | Per capita income |
| 50 | poor | Percent of county living at or below 185% of the national poverty threshold |
| 51 | vacant | Percent of households in county that are vacant |
| 52 | mobilehomes | Percent of households in county that are mobile homes |
| 53 | ageunit5 | Percent of households in county that are <5 years old |
| 54 | avgageunit | Average age of household in county |
| 55 | mediangrossrent | Median gross rent |
| 56 | nocashrent | Percent of households in county where no cash rent is paid |
| 57 | medianhvalue | Median home value |
| 58 | latitude | Latitude of county centroid |
| 59 | longitude | Longitude of county centroid |
| 60 | lnpop | Natural logarithm of county population |
| 61 | lnland | Natural logarithm of county land area |
| 62 | lnpoppersqm | Natural logarithm of population per square mile |
| 63 | gwagegap | Average male earnings - average female earnings |
| 64 | sepdiv | Percent of county that is currently separated or divorced |
| 65 | urban | Percent of county living in urban areas or urban clusters |
| 66 | homeless | Percent of county that is homeless |
| 67 | housingincomediscrep | pci/medianhvalue |
| 68 | over55 | Percent of county that is over 55 |
| 69 | age18to35 | Percent of county aged 18-35 |
| 70 | european | Percent of county born in Europe |
| 71 | asian | Percent of county born in Asia |
| 72 | latinamerican | Percent of county born in Latin America |
| 73 | boomers | Percent of county aged 46-54. 'boomers' is inappropriate, by 10 years...forgot to carry the 1! |
| 74 | bankspe1000pop | Number of banks in the county per 1000 population |
| 75 | vcper1000 | Violent crime per 1000 population |
| 76 | grad | Percent of county whose highest educational attainment is a graduate or professional degree |
| 77 | voterparticip | Percent of county that voted in 2000 |

# References


[1] U.S. Census Bureau, ed. *U.S. Counties Data File Downloads.* URL: `http://www.census.gov/support/USACdataDownloads.html`.