16th Annual Celebration of Student Research and Creative Endeavors

Using Statistical Learning to Predict North Carolina County Voting Patterns

Will Callaway, Eitan Lees, Maureen O'Donnell

Department of Mathematical Sciences

Appalachian State University, Boone, North Carolina

Overview

- On May 8, 2012, North Carolina voters approved Amendment One. This poster examines four different models used to predict North Carolina county voting behavior.
- To ensure accurate predictive power for future observations, the data are split into a training set (80%) and a test set (20%).
- Root mean squared error of the test set is used as a measure of model adequacy.
- All computations and graphs are created with the open source software R [1].

K-Fold Cross-Validation

- Cross validation is the simplest and most widely used method for estimating prediction error [2]. This method directly estimates the expected extra-sample error, $Err = E[L(Y, \hat{f}(X))]$. In this work, the loss function, L, is the square root of the average squared error loss.

- The data in this project is split into $K = 10$ equal sized parts. The cross-validation estimate of the prediction error is

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-K(i)}(x_i)),$$

where $\hat{f}^{-K}(x)$ denotes the fitted function with the K^{th} part of the data removed.

Basic Models Used

I. Least Squares Regression

Note: Variables are described in the Variable Table handout.

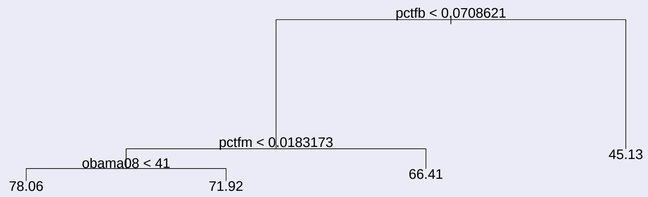
a. Model from [3] applied to Training data (mod1A) — Percent voting for Amendment One is modeled using the predictors pct18.24, medinc, pctb, mccain08, evanrate, pctrural, and pctba.

b. Our OLS model applied to Training data (mod1B) — Percent voting for Amendment One is modeled using the predictors obama08, pctrural, pctw, pctd, pctb, log(pct18.24), log(pctcolenrol), pctfm, log(pctfd), pctown, medinc, medinc^2, evanrate, evanrate^2, pctfb, pctfb^2, log(pctstud), and log(colden).

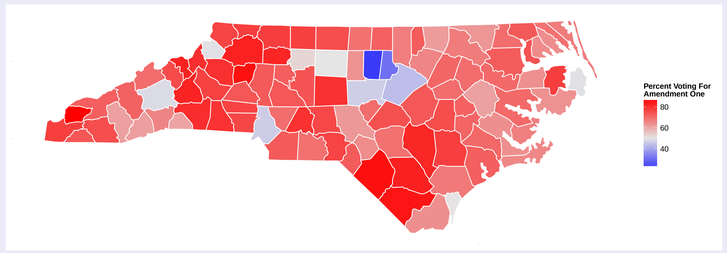
II. Cross validated, $K = 10$, and pruned, $n_{\text{leaves}} = 4$, regression tree [2] (mod2)

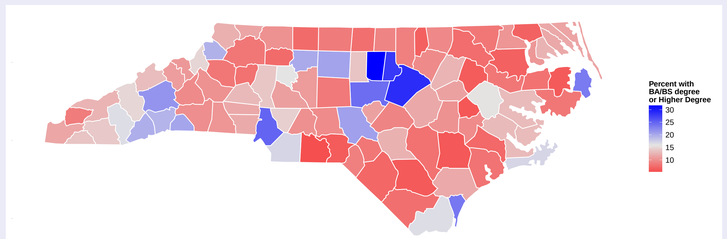
III. Random Forest built from, $n_{\text{trees}} = 5000$, [4] (mod3)

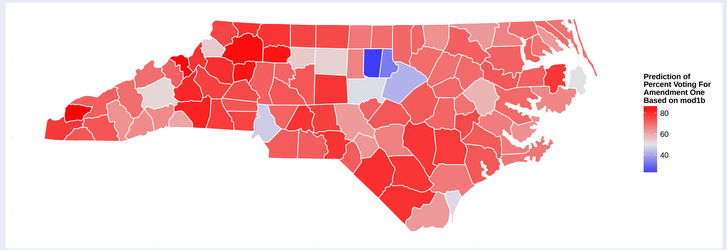
Cross Validated and Pruned Tree



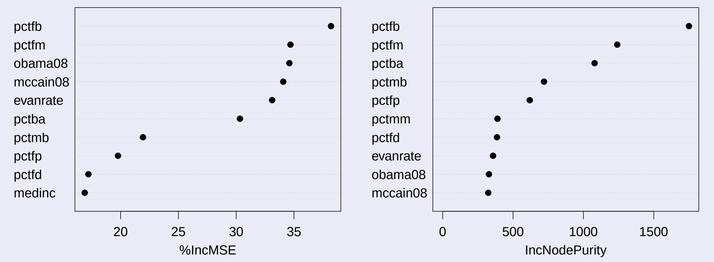
North Carolina Maps







Random Forest Variable Importance



Prediction Errors

	mod1A	mod1B	mod2	mod3
Training Error	3.56	3.05	5.51	5.48
Testing Error	4.95	3.39	7.42	5.71

Table: Training and Testing Error are the RMSE computed from a $K = 10$ cross-validated model for all models except mod3. The RMSE for the random forest model (mod3) does not use cross-validation.

Further Directions

- Use the models developed in this poster to predict county votes for states that have had similar marriage amendments such as South Carolina, Wisconsin, South Dakota, Florida, Idaho, Alabama, Utah, Michigan, Texas, Arkansas, Louisiana, Kansas, Kentucky, Ohio, and Nebraska.
- Use ensemble methods (combining multiple models) for better prediction.
- Make our local maps available via the internet using the shiny server.

References

[1] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, http://www.R-project.org.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference , and Prediction*, New York, New York, 2009.

[3] Davison, E.L. and Jessica N. Eatman *An Ecological Examination of North Carolina's Amendment One Vote to Ban Same Sex Marriage*, (In progress), 2013.

[4] Andy Liaw and Matthew Wiener, *Classification and Regression by* `randomForest`, R News, **2**, 3, 2002, 18-22.

Faculty Advisor: Alan Arnholt

https://github.com/alanarnholt/STT3851-SP2015-Arnholt-Alan