DIMENSIONLESS
TECHNOLOGY

Election Forecasting

Introduction

- We'll be using polling data from the months leading up to a presidential election to predict that election's winner.
- We'll build a logistic regression model.
- Select the variables to include in these models.
- Evaluate the model predictions.

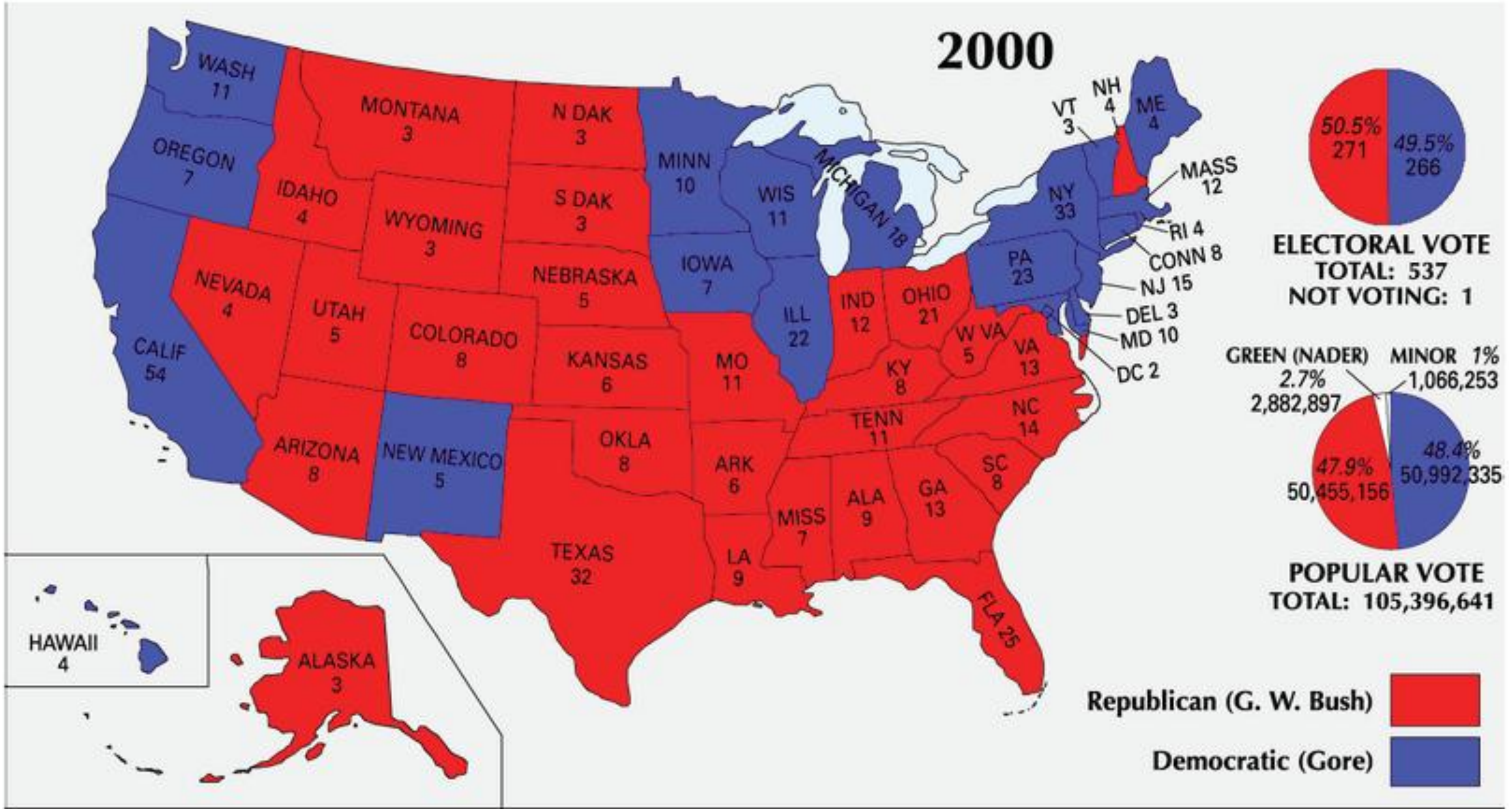
United States Presidential Elections

- A president is elected every four years
- Generally, only two competitive candidates
  - Republican
  - Democratic

The Electoral College

- The United States have 50 states
- Each assigned a number of *electoral votes* based on population
    - Most votes: 55 (California)
    - Least votes: 3 (multiple states)
    - Reassigned periodically based on population change
- Winner takes all: candidate with the most votes in a state gets all its electoral votes
- Candidate with most electoral votes wins election

2000 Election Bush Vs.Gore



Election Prediction

- Goal: Use polling data to predict state winners
- Then-*New York Times* columnist Nate Silver famously took on this task for the 2012 election

The Dataset

- Data from RealClearPolitics.com
- Instances represent a state in a given election
  - *State*: Name of state
  - *Year*: Election year (2004, 2008, 2012)
- Dependent variable
  - *Republican*: 1 if Republican won state, 0 if Democrat won
- Independent variables
  - *Rasmussen*, *SurveyUSA*: Polled R% - Polled D%
  - *DiffCount*: Polls with R winner – Polls with D winner
  - *PropR*: Polls with R winner / # polls

Understanding the Data

- Read the data from the file https://storage.googleapis.com/dimensionless/Analytics/PollingData.csv
- Save it in dataframe “polling”
- Look at the structure of “polling”.
- You might have noticed even though there are 50 states and three election years, so we would expect 150 observations, but we actually only have 145 observations in the data frame.
- Find out the missing observations ?
- We will be making predictions just for the states which have data for all the 3 years.

Cleaning the Data

- You might have noticed that there are these NA values, which signify missing data.

- Find out how many values are missing and for which variable.

- There are various approaches to take for missing data. Read about them in the next slide.

Simple approaches to missing data

- Delete the missing observations
  - We would be throwing away more than 50% of the data
  - We want to predict for all states
- Delete variables with missing values
  - We want to retain data from Rasmussen/SurveyUSA
- Fill missing data points with average values
  - The average value for a poll will be close to 0 (tie between Democrat and Republican)
  - If other polls in a state favor one candidate, the missing one probably would have, too

Multiple Imputation

- Fill in missing values based on non-missing values
  - If Rasmussen is very negative, then a missing SurveyUSA value will likely be negative
  - Just like *sample.split*, results will differ between runs unless you fix the random seed
- Although the method is complicated, we can use it easily through R's libraries
- We will use Multiple Imputation by Chained Equations (mice) package

MICE

•Install and then load a new package, the "mice" package.

•For our multiple imputation to be useful, we have to be able to find out the values of our missing variables without using the outcome of Republican.

•So we're going to create a new data frame called “simple”, which is limited to just the four polling related variables ie. Rasmussen,SurveyUSA, PropR, and DiffCount.

• Look at the summary of “simple” dataframe. We still have our missing values. All that's changed is now we have a smaller number of variables in total.

•To make sure that everybody following along gets the same results from imputation, we're going to set the random seed to the value “144”.

Imputation

- Create a new data frame called “imputed”, by using the function “complete”, called on the function “mice”, called on “simple”.
- Output will show that five rounds of imputation have been run, and now all of the variables have been filled in.
- Look at the summary of “Imputed”. Rasmussen and SurveyUSA both have no more of those NA or missing values.
- The last step in this imputation process is to copy the Rasmussen and SurveyUSA variables back into our original “polling” data frame, which has all the variables for the problem.

Building Models

Building Model

- Split the data into train set and test set
- We're going to train on data from the 2004 and 2008 elections.
- We're going to test on data from the 2012 presidential election.
- Create a data frame called "Train", for train data and "Test" for test data
- What is the baseline model for our Train data? What is its accuracy?
- You will find out that accuracy is only 53%.It always predicts Republican, even for a very landslide Democratic state, where the Democrat was polling by 15% or 20% ahead of the Republican.

Sophisticated Baseline Model

- A reasonable smart baseline would be to just take one of the polls-- in our case, we'll take Rasmussen-- and make a prediction based on who poll said was winning in the state.

- So for instance, if the Republican is polling ahead, the Rasmussen smart baseline would just pick the Republican to be the winner. If the Democrat was ahead, it would pick the Democrat. And if they were tied, the model **would not** know which one to select.

Computing the Baseline Model

- We're going to use a new function called the "sign" function.
- If it's passed a positive number, it returns the value 1.If it's passed a negative number, it returns negative 1.And if it's passed 0, it returns 0.
- So if we passed the Rasmussen variable into sign, whenever the Republican was winning the state, meaning Rasmussen is positive, it's going to return a 1.
- So for instance, if the value 20 is passed, meaning the Republican is polling 20 ahead,it returns 1.
- 1 signifies that the Republican is predicted to win.

Computing the Baseline Model

- If the Democrat is leading in the Rasmussen poll, it'll take on a negative value.
- So if we took for instance the sign of -10, we get -1.
- So -1 means this smart baseline is predicting that the Democrat won the state.
- And finally, if we took the sign of 0,meaning that the Rasmussen poll had a tie, it returns 0, saying that the model is inconclusive about who's going to win the state.

Computing the Baseline Model

- Compute the prediction for all of our training set by applying “sign” function on Rasmussen data.
- Look at the breakdown of our predicted values.
- The smart baseline predicted 56 wins for Republican, 42 for Democrats and 2 inconclusive.

Comparing the Baseline prediction with actual results

- Compare the smart baseline predictions with actual results of train data.
- Calculate the accuracy for the smart baseline model.

Logistic Model

- Consider the possibility that there is multi-collinearity within the independent variables, and there's a good reason to suspect so, as all the variables are measuring the same thing i.e. how strong the Republican candidate is performing in the particular state.
- Compute the correlation between variables.
- You will find, high values of cor-coeff.
- Eg:- SurveyUSA and Rasmussen has a r value of 0.94

Logistic Model

- We will start by building the model by one variable only.
- The one that is most highly correlated with the outcome, Republican.
- We will call this model modl.
- Look at the summary of the output and find the AIC value.
- Let's see how this model does in making predictions.

Making Predictions

- Make predictions on the training set using the modl model.
- We'll call it predl.
- See the outcome measures of the prediction with threshold=0.5
- The accuracy of this model is very close to the smart baseline model.
- We need to improve our model.

Two Variable Model

- We need to make a model with 2 variables.
- Choose 2 variables which have the least multi-collinearity.
- We'll call it mod2.
- Make predictions on the training set using this model.
- See the outcome measures of the prediction with threshold=0.5.

Two Variable Model

- You will find same level of accuracy.
- But AIC value is less, so this is a better model.
- However, neither of these variables has a significance of a star or better.
- So there are definitely some strengths and weaknesses between the two-variable and the one-variable model.

Making Predictions

- Firstly, we will use smart baseline model to make predictions on test data.
- The predictions will be simply the "sign" function applied on "Rasmussen" variable of test data.
- Find the accuracy of this model on test data.

Making Predictions

- Now make predictions using mod2 model.
- We will call it TestPrediction
- Calculate the outcome measures for these predictions for t=0.5
- There is no sense in using the ROC curve for various thresholds, as there is no priority of one type of error over another.
- We are only concerned about accuracy.

Error Analysis

- Pull out the error we have made using sub-setting.
- You can see all the other polls are predicting the Republican win for the State of Florida.
- Overall, the model is outperforming the smart baseline model.