2016 Election: Revisited
Final Report
STAT 545

Travis Benedict and Jordan Pflum

December 12, 2018

Contents

1 Overview

Over two years later, the 2016 Presidential Election remains a frequent topic of conversation. Donald Trump′s victory over Hillary Clinton came as a surprise to many Americans, and as such, there has been a lot academic research into trying to understand the demographic patterns that led voters in key states to choose Trump over Clinton. Our project hopes to continue this academic research by analyze these demographic patterns and seeing if any accurate predictions can be made.

2 Project Background

2.1 2016 Election Background

The 2016 United States presidential election, held on November 8, 2016, was the 58th American presidential election. Despite losing the popular vote by 2.8 million votes, the Republican nominee, Donald Trump, defeated the Democratic nominee Hillary Clinton, winning 30 states and collecting a decisive 304 electoral votes. The messages of the two candidates were fairly different from one another throughout the campaign. Trump′s populist, nationalist campaign, promising to "Make America Great Again" starkly contrasted Clintons expansion and promotion of racial, LBTQ, and women′s rights. Similarly, their target audiences were quite distinct from one another, with Trump targeting the white working-class voters outside major cities while Clinton attempted to appeal to women, hispanic, and younger voters.

2.2 Political Statistics

Many times in politics and elections, it is much more useful to determine to what degree a county, precinct, or state will vote Democrat or Republican, not merely whether it will be won by Democrats or Republicans. For example, if a county is overwhelming Republican, there is little point for Democrats to waste resources in that county when their resources could be put to far better use winning a Toss-Up county. Essentially, parties only care about counties that they have a chance at winning.

Thus, when using statistics in politics, it is advantageous for parties to predict the partisan lean of a county. Partisan lean is a method of binning election results into ordinal categories based on the percent margin of victory for the winning party. Although there are many different ways of characterizing these leans, we have chosen one which splits political leans into nine categories as shown below.

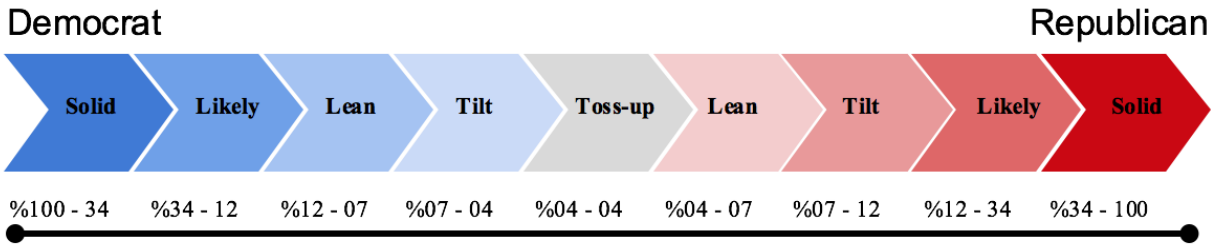


Figure 1: Political Lean Spectrum

The interpretation of the leans is rather straightforward. Say 65 percent of voters in a county voted for Trump while the remaining 35 percent voted for Hillary. Trump would win the election by a margin of 15 percent, and thus, the county would be characterized as "Likely Republican".

Applying the lean categories to the 2016 results by county returns a map which depicts the political lean of counties across America (See Figure 1 for partisan lean descriptions).

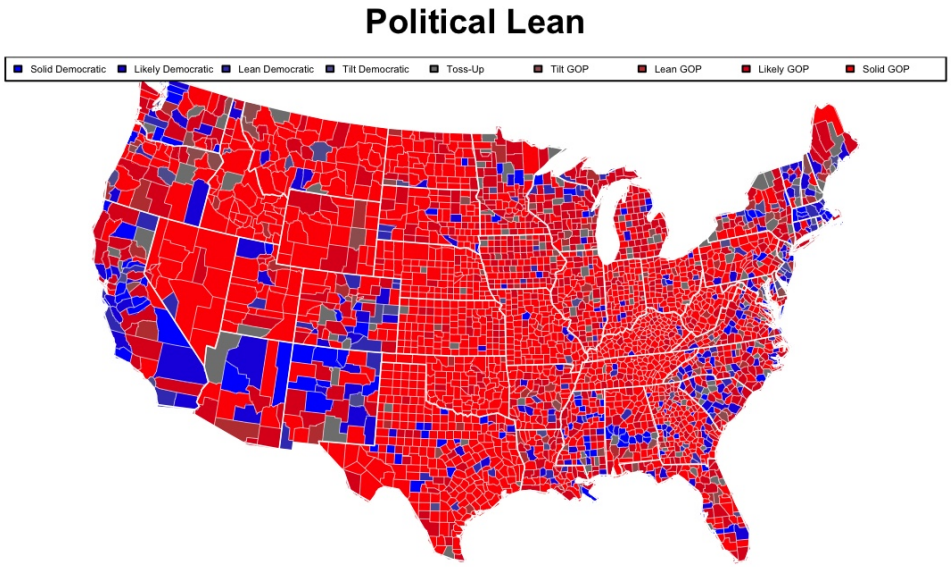


Figure 2: Political Lean of America by County

It may seem puzzling that an election could be so close in light of the fact that an overwhelming amount of counties lean Republican. However, what one has to remember is that many of the counties across the US are sparsely populated while others are heavily populated. For example, Harris County contains 4.653 million residents while Loving County contains only 134 residents. Additionally, historically, cities tend to lean Democratic while rural areas tend to lean Republican.

3 Data Overview

3.1 Data Set Desccriptions

We acquired our datasets from Kaggle on a page called "2016 US Election". We had three data sets in total. Our first dataset was our fact data set. It contained more than three thousand American counties as well as 54 demographic predictors for each county. Our second dataset was our election results dataset. It contained the same three thousand American counties as the first dataset but contained election results for each county rather than demographic information. The results included the number of votes for each candidate as well as the percentage of votes for each party. We converted these results into political leans and appended a column to the data set with the calculated political leans. For reference, we assigned the ordinal number one to correspond with strongly Democrat all the way to number 9 which represented strongly Republican. Our final data set was simply a train and test dataset used when running regressions. We split 75 percent of the data into the training set and the remaining 25 percent of the data into the testing set.

3.2 Preditor Examples

The 54 predictors of demographic information for each county covered a wide range of subjects. Since it would be impractical to list all 54 predictors, we took only a small subset to display in this report (shown below).

Predictor Code	Predictor Description
RHI725214	Hispanic or Latino, percent, 2014
POP815213	Language other than English spoken at home, pct age 5+, 2009-2013
EDU685213	Bachelor′s degree or higher, percent of persons age 25+, 2009-2013
INC910213	Per capita money income in past 12 months (2013 dollars), 2009-2013
PVY020213	Persons below poverty level, percent, 2009-2013
PST045214	Population, 2014 estimate
AGE295214	Persons under 18 years, percent, 2014
AGE775214	Persons 65 years and over, percent, 2014
SEX255214	Female persons, percent, 2014
RHI225214	Black or African American alone, percent, 2014
VET605213	Veterans, 2009-2013

3.3 Data Cleansing

As always, some data cleansing was required. Fortunately, our data set had no missing values. However, in addition to county data, our data set also included statewide data. Since the county data already captured and information contained in the statewide data, we decided to remove the statewide data from the data set. Additionally, there were a few discrepancies in the unit of measurements between the predictors. Knowing this, we decided to scale the data to avoid any problems with variable selection or regressions.

4 Exploratory Analysis

4.1 Overview

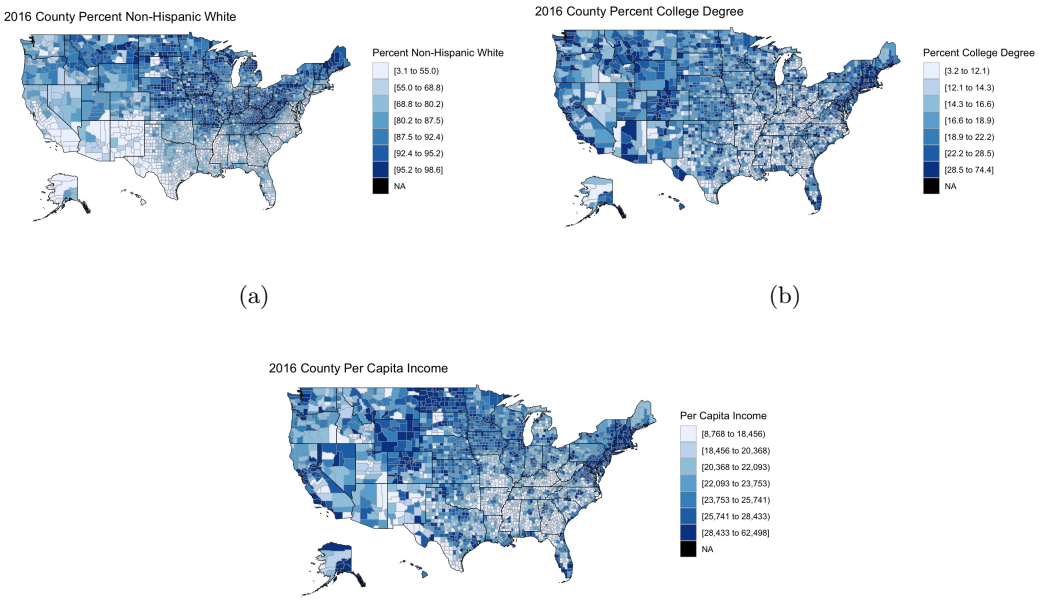
Before diving into the project, we wanted to conduct an exploratory analysis to familiarize ourselves with the data as well as to see if there were any interesting trends we wished to analyze further.

A major theme throughout the election was the importance that Trump′s base played in his victory. Remembering this, we decided it would be interesting to find if the data reflected this.

Trumps base was characterized as white working-class voters outside major cities. We thought three predictors which would accurately capture his base were the percent of non-hispanic white people, the percent of people that have obtained a college degree, and the per capita income for each county.

4.2 Maps

A logical way to analyze the important characteristics of Trump′s base is by looking at their geographical distribution. Thusly, we mapped each characteristic onto a map of the United States divided into counties. Throughout the analysis of the map, we will reference back to Figure 2 which depicts each county′s lean on a map.



(a) (b)

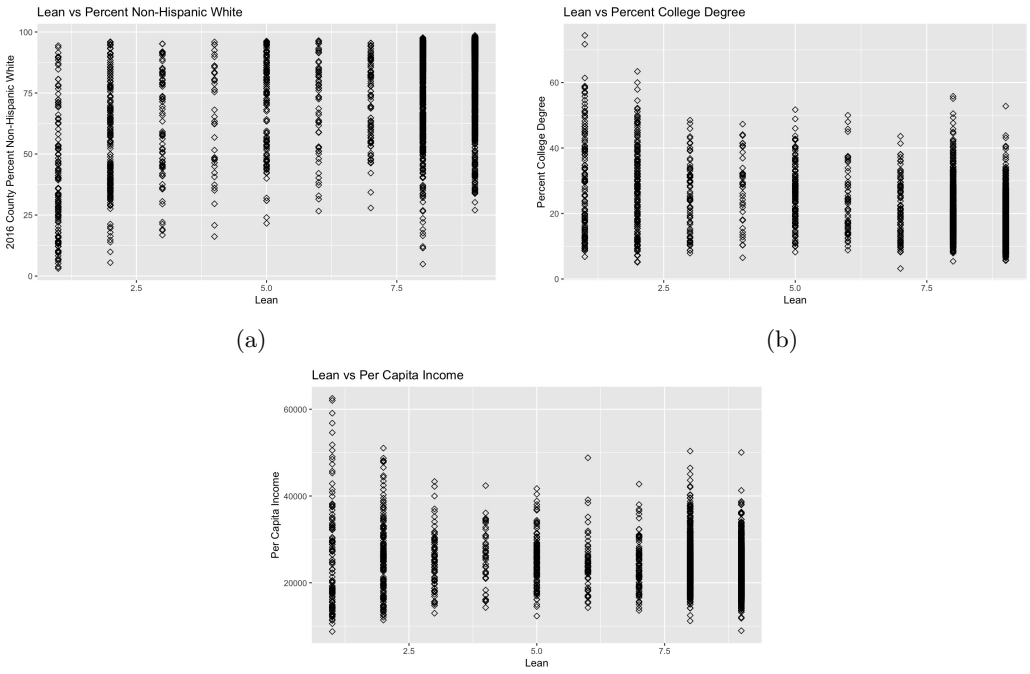
(c)

Figure 3: Trump′s Base Demographics (Maps)

There are a few takeaways from Figure 3.a. First, when comparing Figure 3.a to Figure 2, one can clearly see that many of the areas shaded dark blue (representing to a high percentage of non-hispanic white counties) are located in central America, which corresponds to the highly shaded red areas seen in Figure 2, specifically in central America. However, something our team found interesting is analysing the border coutnies. One would expect border counties, which have a relatively low percentage of non-hispanic white people (as shown in Figure 3.a) to swing left. Yet when you look back to Figure 2, we see that many of the border counties swing right. This tells us that while it would appear that the percent of non-hispanic white people in a county play a role in the partisan lean of a county, it is not the only factor that should be considered. The same type of analysis can be applied to Figures 3.b and 3.c. Both of these maps indicate a high percentage of people with college degrees and a high per capita income in the counties located in the North East. Referencing Figure 2 we see that these are counties that Trump tended to lose, signaling to us that these are important predictors.

4.3 Scatter Plots and Histograms

Although the geographical maps provide insight on the locations of Trump′s base, it may be more useful to represent the data in a more succinent and easier to interpret fashion. An easy way to do this is to plot the key characteristics of Trump′s base onto a scatter plot, analyzing the distribution compared to the political lean.



(a) (b)

(c)

Figure 4: Trump′s Base Demographics (Scatter Plots)

Figure 4.a maps each counties percent non-hispanic white population against its political lean. Figure 4.a clearly shows an upward trend in the data. This trend can be interpreted as follows: as a county′s percentage of non-hispanic white people increase, the county′s political lean tends to become increasingly Republican. Another interesting component of Figure 4.a is the variance of the data for each political lean. Looking at only the extremes, we see that strongly Democratic counties can contain both a large proportion of non-hispanic white people and a low proportion. However, on the other side of the spectrum, counties which are strongly Republican never have a percentage of non-hispanic white population lower than 25 percent, with most of the counties containing a very high percentage of non-hispanic white population. Similar analysis can be applied to Figure 4.b, which maps each counties percent of population holding a college degree against its political lean, and Figure 4.c, which maps each counties per capita income against its political lean. Unlike Figure 4.a, Figure 4.b shows a slight downward trend in the data. This trend can be interpreted as follows: as a county′s percentage of the population which have obtained a college degree decreases, the county′s political lean tends to become increasingly Republican. When analyzing Figure 4.c, it is quite simple to see the increasing variance of the data as counties lean Democratic. This can be interpreted as follows: counties which lean Democratic are more likely to have high per capita income or low per capita income, while counties which lean Republican will most likely have a lower per capita income. All of these observations reenforce the importance of Trump′s base in determining the political lean of a county.

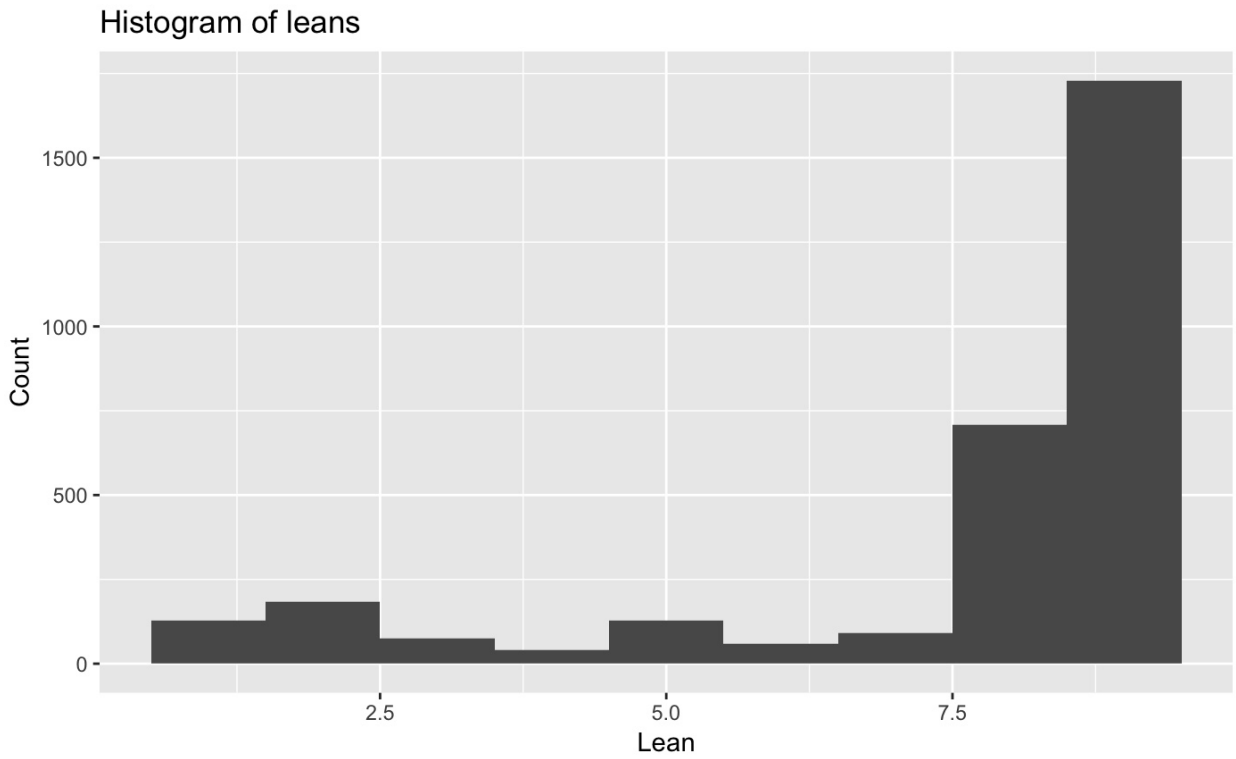


Figure 5: Partisan Lean Histogram

Recalling the political lean map introduced in section 2, one can clearly see that a vast majority of the counties lean Republican. This observation is emphasized in Figure 5, which reveals that not only are a bulk of counties Republican, but an overwhelming amount are categorized as strongly GOP. The histogram reveals that the data is quite unbalanced, which leads to some problems addressed later in this report.

5 Project Goal

After performing the exploratory analysis, we narrowed our focus and developed a central question we wanted answered; does the probability associated with logistic regression outcome provide insight for ordinal regression?

Unpacking this question may be helpful. Logistic regression returns the probability that a county will be either Democrat or Republican. However, a more interesting question to analyze is how this probability relates to the lean probabilities provided by ordinal regression. For example, if logistic regression determines that the probability that a county will be Republican is 75 percent, does that correspond to that county being designated as strongly Republican or some other political lean? If there is some kind of relationship between logistic regression probability and ordinal regression outcomes then we expect to find that variables selected for predicting accurate logistic regression will also lead to accurate ordinal outcomes.

6 Variable Selection

6.1 Overview

After preparing our data we applied several different variable selection methods in hopes of eventually fitting an accurate logistic regression model for predicting if Hillary Clinton or Donald Trump would win a county. Variable selection allows us to explain the data in the simplest way (ie Occam′s Razor) while also reducing the noise that unnecessary predictors add.

6.2 Principal Component Analysis

The first method we performed was a Principal Component Analysis on the training data. By applying PCA we were able to transform our data into a set of orthogonal column vectors, sorted in decreasing order of the proportion of variance each column explains. As demonstrated by the Scree plots below, the first principal component of of the data accounts for about a third of the total variance with the proportion of variance explained by each subsequent component decreasing relatively quickly. For our model fitting we decided to test the efficacy of the first 5 and first 10 principal components; these accounted for roughly 66 and 80 percent of the variance, respectively.

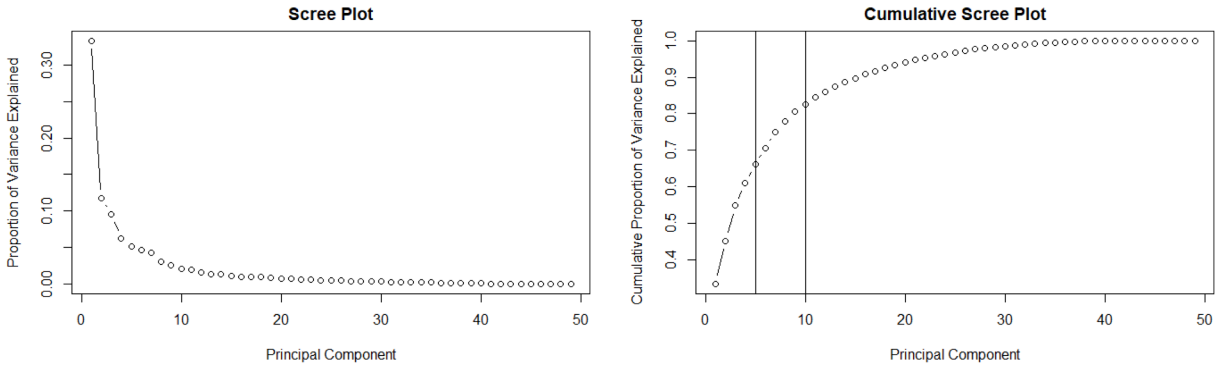


Figure 6: Variance Captured by PCA

6.3 Forward/Backward AIC

The second method we applied was Forward and Backward Stepwise Selection Algorithms for Akaike Information Criterion (AIC). AIC provides a penalized maximum likelihood fit that is meant to decrease overfitting. The formula is as follows:

$$AIC = -2\text{loglikelihood} + 2k = N \times log\left(\frac{RSS}{N}\right) + 2k$$

where N is the number of observations, RSS is the residual sums-of-square, and k is the number of free parameters. Although larger models will fit the data better, and therefore have a smaller RSS, they will also have a large number of parameters. Thus, our goal is to balance the fit of the model with its size. Generally, the smaller the AIC, the better the model.

6.4 Forward/Backward BIC

The third method we applied was Forward and Backward Stepwise Selection Algorithms for Bayesian Information Criterion (BIC). The formula is as follows:

$$BIC = -2\text{loglikelihood} + k \times log\,(N) = N \times log\left(\frac{RSS}{N}\right) + k \times log\,(N)$$

where N is the number of observations, RSS is the residual sums-of-square, and k is the number of free parameters. BIC behaves very similarly to AIC. However, BIC penalizes larger models more heavily than AIC, so it tends to favor smaller models when compared to AIC.

6.5 Lasso

The final method we applied was Least Absolute Shrinkage and Selection Operator (LASSO). LASSO adds an L1 norm penalty $\lambda\|\beta\|_1$ to standard logisitc regression likelihood function, where λ is a hyperparameter determined by a grid search using the R package $'$glmnet$'$ and β is the vector of regression coefficients.

7 Application and Comparison

After applying the five non-PCA methods we examined which variables were selected most frequently, hypothesizing that these variables were the most "important" for making predictions in the logistic regression context. With this idea in mind we fit two different models: a "consensus" model using only variables selected by all five techniques and a "majority" model using variables selected by at least four of the five techniques. The consensus model was based on four variables: percentage of people over 25 with less than a bachelors degree, median value of owner-occupied housing units, median household income and percentage of women owned firms. Among these percentage of women owned firms was the most surprising to us, but intuitively the variable seems like a proxy for measuring involvement of women in politics since business owners tend to have more at stake when it comes to how laws are written.

To measure and compare the success of these variable selection methods we used Receiver Operating Characteristic curves and the associated area under curve value (AUC). The ROC curves provide a visualization of true positive rate compared to true negative rate as a function of the decision threshold for a model. AUC represents the probability that a randomly chosen positive outcome be ranked higher than a randomly chosen negative outcome. An AUC of 1 implies that the binary classifier correctly matched all test examples. The plot below shows ROC curves and associated AUC values for each variable selection technique.

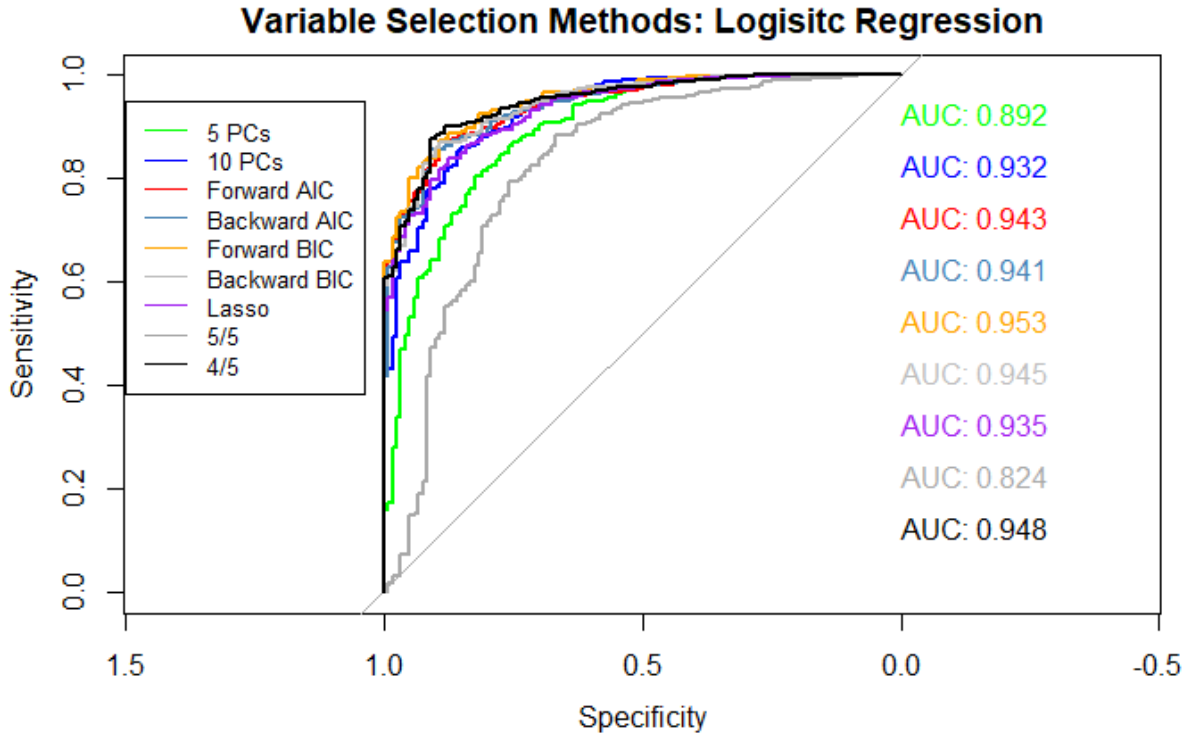


Figure 7: ROC Curves for Logistic Regression

With the exception of the consensus model and the first five principal components model, every variable selection technique performed about the same. Each of these technique were able to select models with AUCs over 0.93, demonstrating that there is enough information in this data set to predict election outcomes with high accuracy. Given this we decided to apply the same set of variables to try to predict the partisan lean for each county in our test set. Treating partisan lean as an ordinal variable from 1 to 9 (Solid Democrat to Solid Republican), we fit a Cumulative Logit Model to each set of variables selected for the logistic case. Below is the ROC curve and associated AUC for each model.

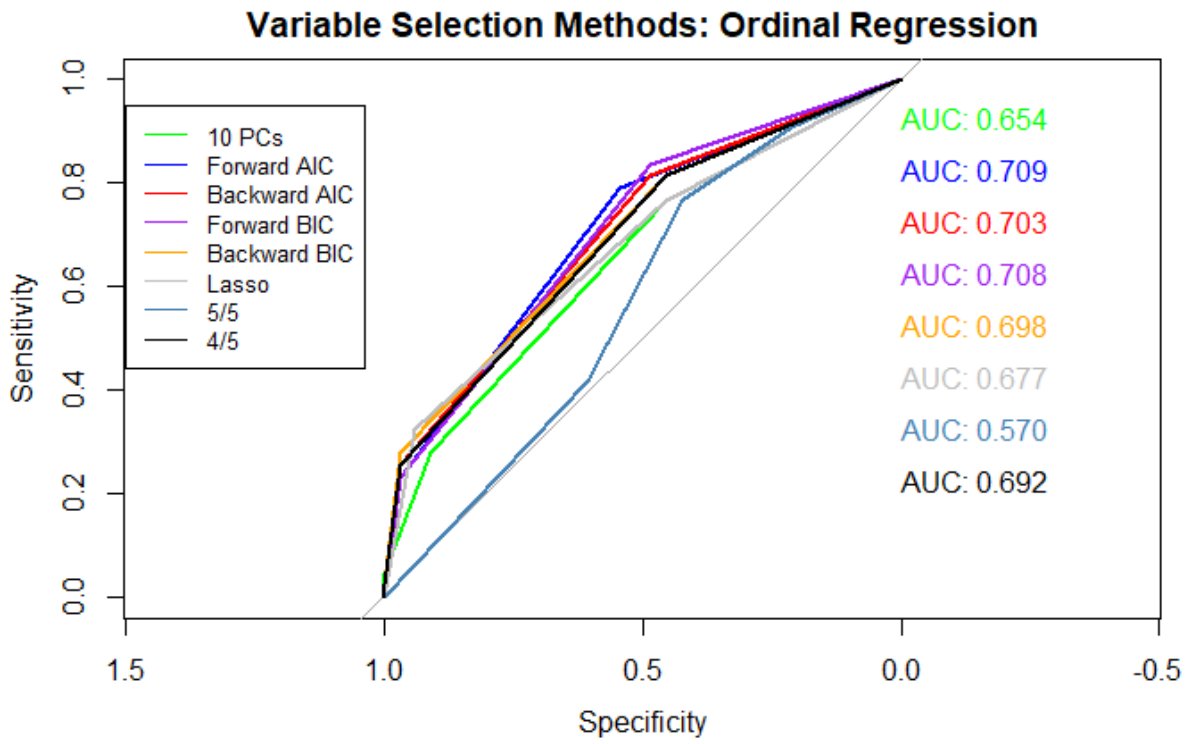


Figure 8: ROC Curves for Ordinal Regression

Clearly the variables selected do not offer as much insight into the partisan lean of each county as they do for predicting the winner of an election. Despite performing about as well as the other models in the binary outcome setting, the principal components model performs noticeably worse than the other variable selection techniques. We hypothesize that since the first ten principal components only account for 80 percent of the overall variance in the data and perform well for predicting the winner of an election then the remaining 20 percent of the variance must better account for the variability between leans. The stepwise methods for both AIC and BIC perform about as well as each other even though BIC models are about half the size of the AIC. To get a better understanding of model performance the confusion matrix for the best model (Forward AIC) is included below with the true label along the top and the predicted label on the side.

	Solid Democratic	Likely Democratic	Lean Democratic	Tilt Democratic	Toss-Up	Tilt GOP	Lean GOP	Likely GOP	Solid GOP
1	18	9	1	0	0	0	0	2	0
2	14	24	4	4	6	2	7	13	0
8	1	9	16	4	26	7	18	93	40
9	0	1	0	0	2	1	4	74	385

Figure 9: Confusion Matrix for Forward AIC Ordinal Regression

TThe model does fairly well at predicting Solid and Likely leans, either getting them correct or misclassifying them into the adjacent category but misses the inner classes entirely, not predicting a single Lean, Tilt or Toss-Up. This demonstrates just how unbalanced the leans in this dataset are. Since there are so many Solid and Likely GOP counties it makes the probability of a county being anything else very small. Although we viewed the partisan lean of a county as an ordinal variable we decided to treat it as nominal in

an attempt to improve the accuracy of predicting Lean, Tilt and Toss-Up. With this in mind we fit Baseline Categorical Logit Models using the variables selected by our original logistic fitting with Solid GOP as our baseline category. The plotted results are included below.

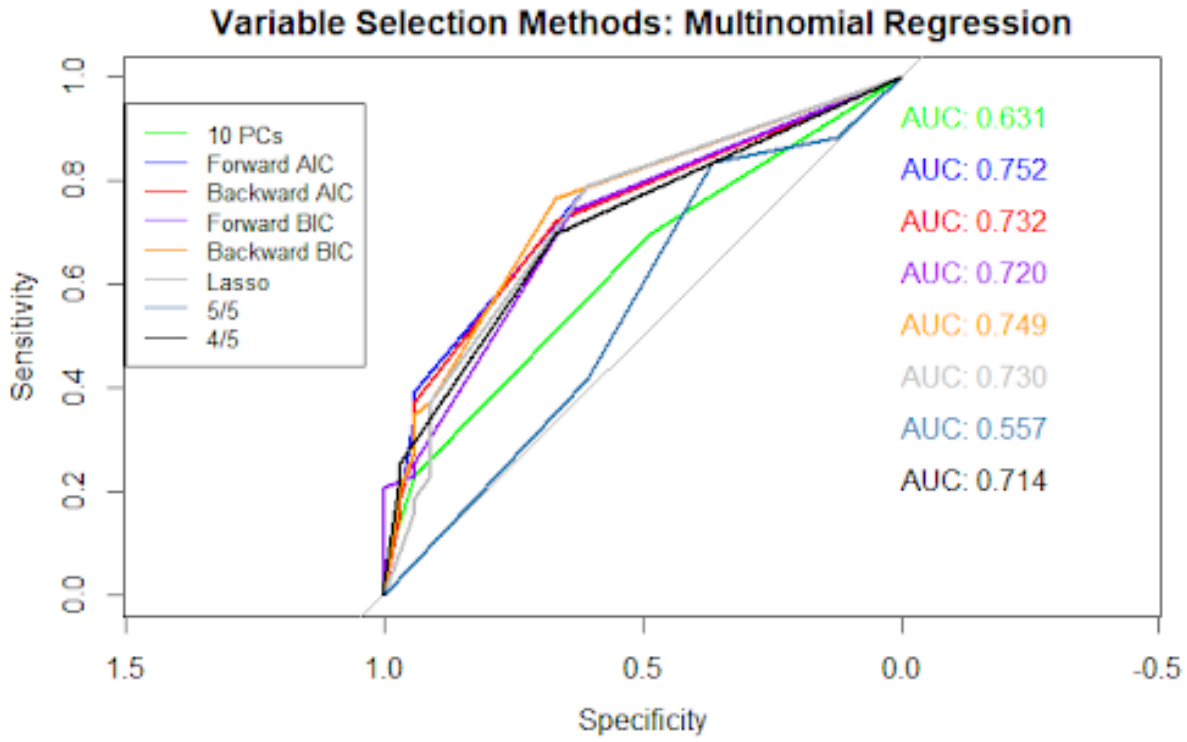


Figure 10: ROC Curves for Multinomial Regression

In general the multinomial models performed noticeably better than the ordinal models. For the most part the analysis about the ordinal models holds here as well. The principal components model does not fare well and the AIC/BIC models perform about as well as each other though there is much more variability between the stepwise models than in the ordinal case. Upon examining the confusion matrix for the best model, Forward AIC, it becomes clear that treating partisan lean as nominal a variable is a far superior approach.

	Solid Democratic	Likely Democratic	Lean Democratic	Tilt Democratic	Toss-Up	Tilt GOP	Lean GOP	Likely GOP	Solid GOP
1	20	6	1	0	1	0	0	1	0
2	9	27	5	4	8	3	1	7	0
3	0	1	0	1	0	0	0	1	0
4	0	0	0	0	2	0	0	0	0
5	4	1	3	1	3	1	1	3	0
7	0	0	0	1	0	0	0	0	0
8	0	6	4	3	25	6	15	80	39
9	0	1	1	0	2	1	3	81	402

Figure 11: Confusion Matrix for Forward AIC Multinomial Regression

Unlike the best ordinal model, the best nominal model actually predicts outcomes in all categories. This

is a significant improvement even if the leans are not always predicted correctly. The ROC curves and AUC values do not fully capture how much better the multinomial approach is since they do not account for the idea that classifying a Toss-Up as Tilt GOP is actually much better than classifying it as a Solid GOP. In this example both are of course incorrect classifications but one is much closer to the truth. Ultimately this represents a shortcoming in our choice of measurement.

8 Conclusion

Based on our analysis we have four main takeaways. First, the conventional wisdom that Donald Trump won the 2016 election because of white, uneducated voters is fair but does not explicitly account for the importance of income level in determining voting patterns. Second, principal component analysis was very successful in predicting binary outcomes using relatively few principal components, but in order to select between multiple classes more principal components are necessary. In terms of determining the best variable selection technique for this type of data it seems that there is little difference in using a stepwise algorithm based on AIC or BIC. Although Forward AIC was the model with the highest AUC in both the ordinal and multinomial models the other stepwise models were essentially as good. Overall the most important conclusion that we drew from this analysis was how much better the multinomial model performed compared to the cumulative logit model despite the data having a fairly straightforward ordinal interpretation.