STAT 541 Term Project: 2016 Election Revisited

Travis Benedict In Collaboration With Jordan Pflum

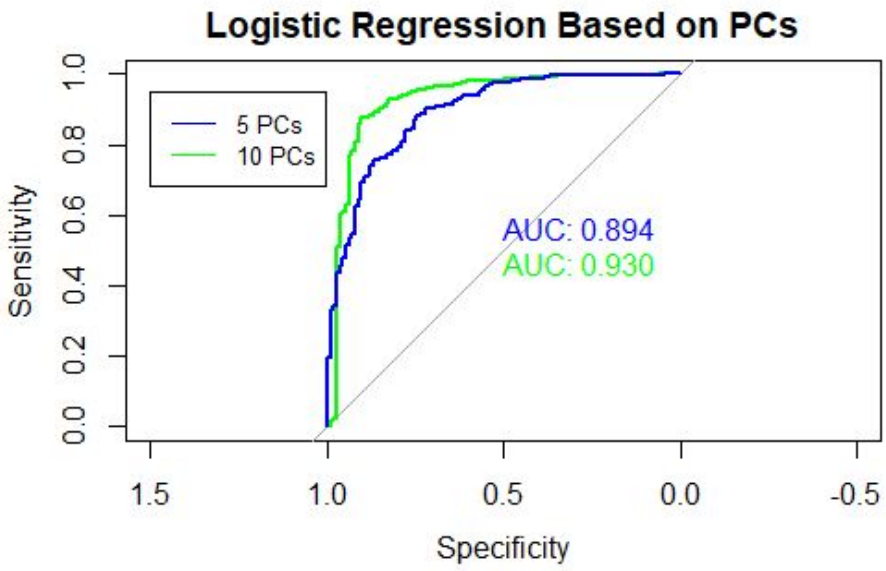
December 11, 2018

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 dataset was chosen with this sort of research in mind. The data, sourced from Kaggle, contains election results for more than three thousand American counties, as well as fifty demographic descriptors for each county[1]. The demographic descriptors include things like total population, percentage of people with at least a bachelor's degree and number of women-owned firms, providing us with a fairly comprehensive understanding of the demographics of each county. Jordan's report describes our exploratory analysis and the data set itself more thoroughly.

Given the wide variety of demographic predictor variables in the data, our goal for this project was to reduce the overall amount of information we had while still gaining reliable insight into voting patterns. With this in mind we began preprocessing of the data to remove statewide information which was also included in the dataset but irrelevant to our countywide analysis. We then standardized the demographic predictor variables to account for discrepancies in units used for measurement. After completing our preprocessing we first applied principal component analysis to the demographic data.

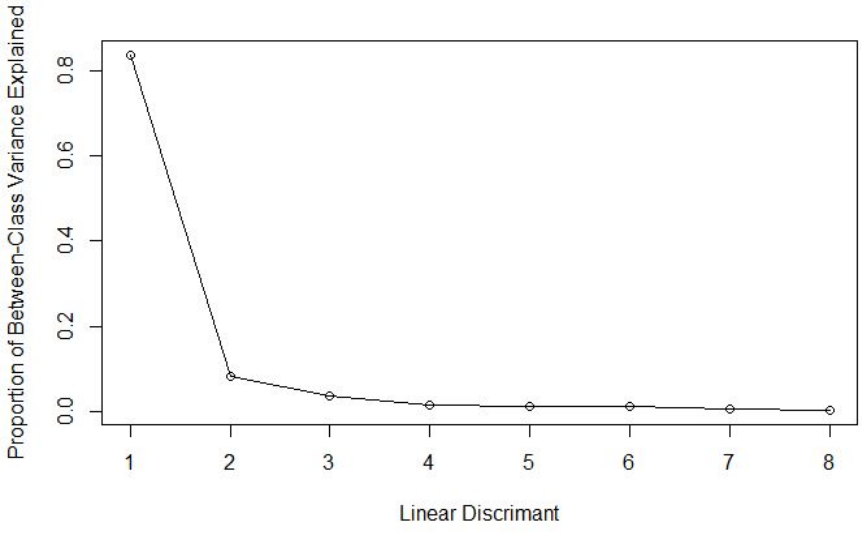
[1] https://www.kaggle.com/benhamner/2016-us-election/home

Using the first five and ten principal components we fit a logistic regression model to determine which candidate won a county. To measure and compare the success of the models 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 is 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 the models based on the first five and ten principal components.



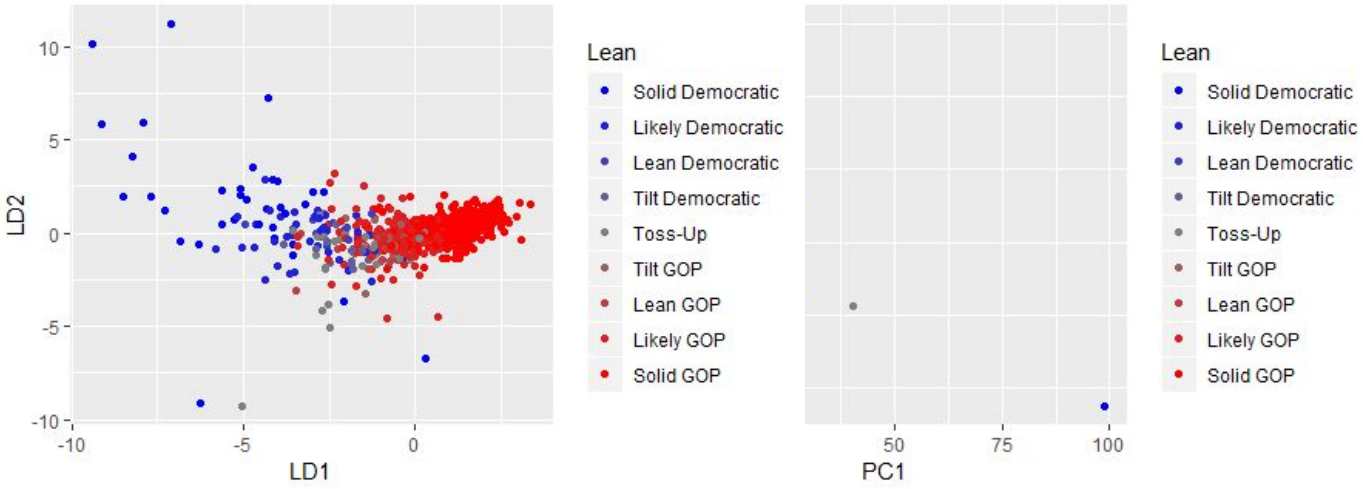
Although the first five and ten principal components only account for 66% and 88% of the total variance of the data, both models performed very well on the test data. For simply predicting the binary outcome of an election it is clear that the dataset we have can be transformed to reduce the dimensionality significantly while still allowing for accurate predictions.

Following the success of Principal Component Analysis we then applied Linear Discriminant Analysis to predict the "partisan lean" of the counties[2]. Partisan lean is a way of binning election results into different categories based on the percent margin of victory for the winning party. In order of increasing competitiveness the leans are Solid, Likely, Lean, Tilt and Toss-Up. For example, a Solid Democrat county represents a county where Clinton won by at least 34 percentage points while a Tilt Gop county means that the margin of victory for Trump was between 4 and 7 percentage points. We chose LDA because it provides a supervised framework for finding a set of continuous independent variables for predicting categorical outcomes. Though the outcome of LDA is similar to that of PCA, we expected that LDA would perform better than PCA because PCA is an unsupervised method and therefore has no understanding of the outcome variable. Each variable output by LDA explains a portion of variance between the output classes, similar to the way that each principal component explains a portion of overall variance. The plot below depicts the proportion of between class variance each linear discriminant explains.



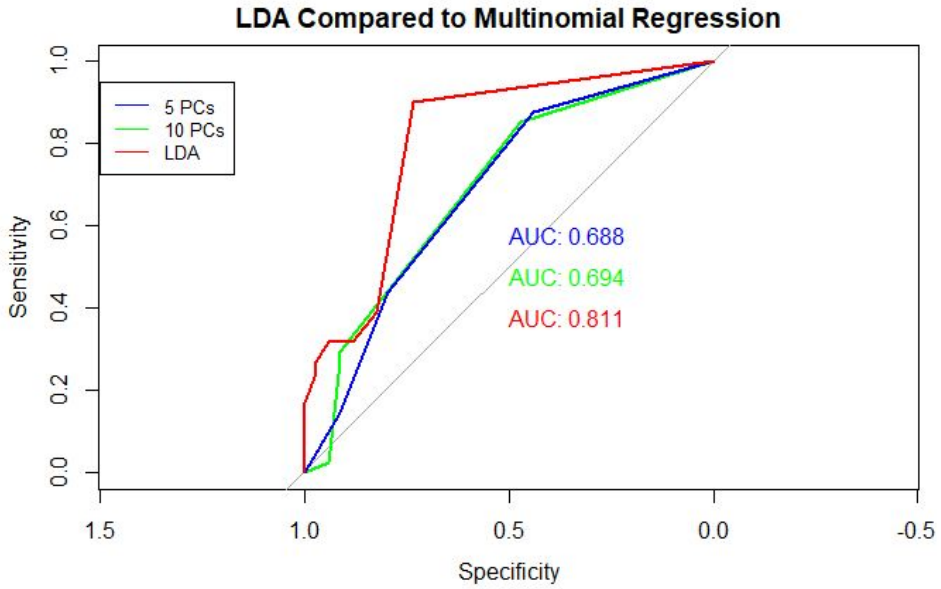
[2] https://fivethirtyeight.com/features/2018-house-forecast-methodology/

The first linear discriminant explains over 80% of between class variance, demonstrating that LDA is a highly effective way of reducing dimensionality for this problem. Given that only eight linear discriminants are needed to explain all of the between class variance we decided to include all the linear discriminants in analysis going forward. This still reduced the dimension of our data from 50 to 8.



Comparing the first two principal components to the first two linear discriminants, as seen above, shows geometrically why LDA is superior to PCA for predicting the partisan leans for a county. Although there appears to be some separation along the first principal component between strong Republican voting counties and strong Democratic ones, the separation is very minimal and does not provide separation for Toss-Ups. By contrast the LDA plot shows that along the first linear discriminant there is fairly well defined separation between strong Democratic and strong Republican voting counties while also providing some separation for Toss-Ups in between the two larger groups.

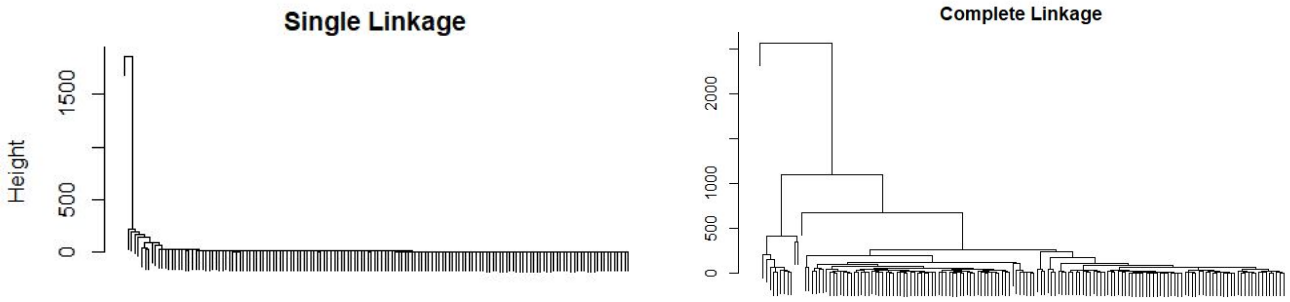
Using the first five and ten principal components, we fit a multinomial regression model to predict the partisan lean for a county. The plot below compares the ROC curves for these two models with the ROC curve for the LDA.



As expected LDA significantly outperforms PCA. Despite the first 10 principal components explaining approximately 15% more of the overall variance than the the first five principal components, the two methods perform about the same. Based on our examination of the confusion matrices for the PCA model, it seems that the unbalanceness of the output classes is an important factor in causing the misclassification. The PCA models failed to predict any of the counties as being Lean, Tilt or Toss-Up; instead they often misclassified these less obvious outcomes as Solid or Likely Republican, the two most frequent outcomes.

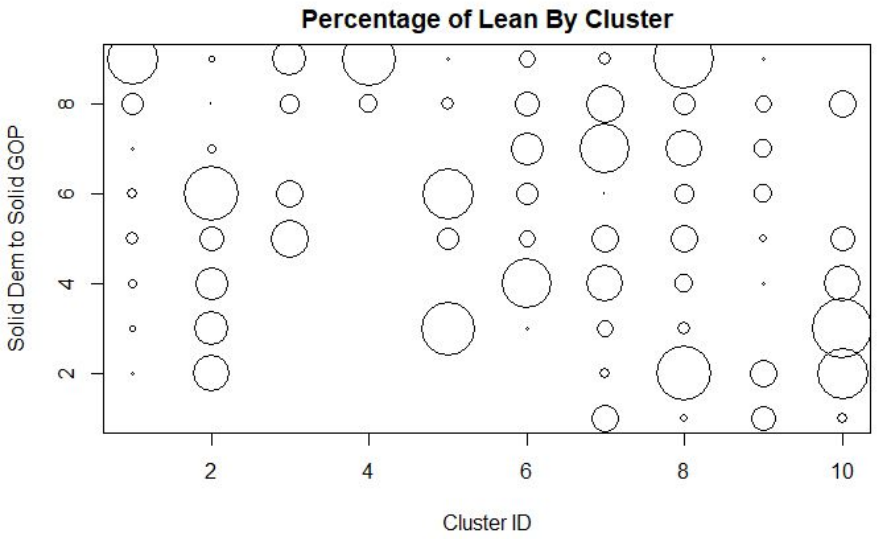
In an attempt to further reduce the dimensionality of our data we next attempted to cluster the counties based on their demographic facts, hoping to find a correlation between clusters and partisan lean. We initially attempted hierarchical clustering using the scaled data, excluding

variables that were found to have a correlation of 0.8 or greater with any other variable. Below are the dendrograms using a random sample of 150 counties from our test data.



As shown by the dendrograms it is clear that single linkage results in chaining, meaning that complete linkage is preferable in this setting. The complete linkage dendrogram shows that the clusters are not especially well defined and may be less informative due to outliers. Clustering based on principal components and linear discriminants resulted in similarly inconclusive results.

Since hierarchical clustering was unable to give us clear results we turned to the EM algorithm and the MClust package in R. Using the "mclustBIC" function we found the optimal number of clusters for several different inputs, ultimately hoping that we would be able to find some kind of correspondence between leans and clusters. Each input that we tried was a different form of our scaled data including: the entire dataset, the variables with a correlation less than 0.8, the first ten principal components and the linear discriminants. Each of these methods was determined to have an optimal number of clusters between 5 and 10, which initially seemed promising to us given the nine possible partisan leans. To get an understanding of how the clustering algorithms performed we plotted Cluster ID vs partisan lean where the size of each point represented the percentage of the lean captured by the point. Below is the plot for clustering based on the linear discriminants.



As shown by the plot most of the clusters are uninformative. Cluster #4 is a good example of what we hoped to see with the clusters; it represents a substantial percentage of all Solid Republican counties and a much smaller percentage of Likely Republican counties. This cluster is well defined and informative. Meanwhile, Cluster #8 represents a wholly uninformative cluster. It contains large percentages of Solid R, Likely R and Likely D along with small percentages of everything else. Across all of our clustering trials our results tended to look more like Cluster #8 rather than Cluster #4. This led us to conclude that for our goal of reducing information while still gaining insight into voting patterns clustering was not an effective solution, especially in comparison to PCA and LDA.

Based on the techniques applied for this analysis we have concluded that the dimensionality of the data can be significantly reduced while still providing predictive power for determining countywide voting patterns. Predicting the winner of an election can be done with high accuracy using just the first ten or even five principal components. LDA is highly effective

for predicting the partisan lean of a county while reducing the data down to just the eight linear discriminants. Unlike PCA and LDA, we determined that clustering was ineffective for gaining insight into county votings. Ultimately, this work does not provide any insight into the 2016 election that is necessarily novel, but it does provide strategies that can be used for modeling future elections.

References

https://www.kaggle.com/benhamner/2016-us-election/home

https://fivethirtyeight.com/features/2018-house-forecast-methodology/