**Capstone Project Proposal**

Sujoy Gayen

December 18, 2017

**Abstract**  
This project will create a model to predict on a county by county basis the percentage of the vote that the Republican and Democratic presidential candidates in the U.S. presidential 2016 election received. This model could be used by political analysts and campaign personnel in future presidential elections to predict the percentage of the vote that each of the two major party candidates would receive.

**Problem Setup**

The U.S. presidential election of 2016 was a surprising historic election. Against all odds, despite the predictions of nearly every political commentator, Donald Trump beat Hillary Clinton. Granted, Hillary Clinton won the popular vote. However, Donald Trump won enough votes in key states to win the electoral college. Many explanations have been put forward for Trump's victory. Many political commentators have noted that Trump appeared to win many traditionally Democratic voters, including voters that had voted for the Democratic president Barack Obama in the previous 2 presidential elections. Specifically, political commentators noted that Trump appeared to win an unusually large proportion of the white working class vote, especially in states such as Pennsylvania, Ohio, Wisconsin, and Michigan that had voted for Obama in the prior 2 elections. This capstone project aims to model the determinants of the 2016 presidential vote on a county by county basis.

**Data**

The data set is available here at this link: <https://github.com/Deleetdk/USA.county.data>

A good portion of this data came from the New York Time’s 2016 presidential election county-by-county-analysis.

The overall dataset includes 13 attributes for 3145 observations.

The observations come from each county in the United States.

The attributes with their corresponding variable names include:

* Fips: 4 digit county identification number
* Name of the county
* Average age of the inhabitants of the county: variable name is average\_age
* Percentage of the population of the county that is white: percent\_white
* Percentage of the population that lacks health insurance: percent\_uninsured
* Percentage of the population that has an educational degree: percent\_degree
* Average income of the inhabitants of the county: average\_income
* State that the county is located in
* Percentage of the population of the county that voted for Republican candidate Donald Trump in 2016 presidential election: percent\_republican
* Percentage of the population of the county that voted for Democrat candidate Hillary Clinton in 2016 presidential election: percent\_democrat
* Percentage of the population of the county that voted for Republican candidate Mitt Romney in 2012 presidential election: percent\_republican\_2012
* Percentage of the population of the county that voted for Democrat candidate Barack Obama in 2012 presidential election: percent\_democrat\_2012

**Analysis Approach**

This dataset lends itself to a linear regression model. I will build a model that predicts the percentage share of the vote in each county that Republican presidential candidate Donald Trump received in the 2016 presidential candidate. I will divide the data into a training set and testing set. The training set will contain 80% of the observations, while the test set will contain twenty percent of the observations. Initially, I will run a linear regression model on the training set using all the explanatory independent variables in the dataset. After creating this initial model, I will remove variables that are not significant (based upon p value and t-test value) and remove variables that have a high degree of collinearity to create a more effective simplified value. My goal in creating a model for this training set is to maximize R^2. After finding an optimal model for the training set, I will run my predictive mode on the test set. I will add a new variable to the test set that consisted of all my model’s predicted values for each observation in the test set. From there, I will calculate the mean standard error and the R2 value to determine how good of a fit my model is for the test set.

**Deliverables**

The primary deliverables are the code used to execute the analysis described above, and the final code which would also include data visualization of my key variables.

I will also deliver a paper and slide deck which will communicate the analysis process and the findings. The paper and slide deck will be targeted to a potential audience of political analysts and campaign personnel who might presumably use the model for predicting outcomes of future presidential elections.

All final deliverables, as well as intermediate deliverables, will be made available on the project's GitHub Repository.